SECURITY CLASSIFICATION OF THIS PAGE

AD-A213 273

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION: Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT: Approved for public release distributed unlimited
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S)	5. MONITORING ORGANIZATION REPORT NUMBER(S): AFOSR-TR-89-1511
6a. NAME OF PERFORMING ORGANIZATION: University of Kansas; 6b. OFFICE SYMBOL (If applicable)	7a. NAME OF MONITORING ORGANIZATION: AFOSR/NL
6c. ADDRESS (City, State and ZIP Code): Department of Psychology, University of Kansas, Lawrence, Kansas 66045	7b. ADDRESS (City, State and ZIP Code): AFOSR/NL, Building 410, Bolling ARB DC 20332-6448
8a. NAME OF FUNDING/SPONSORING ORGANIZATION: AFOSR; 8b. OFFICE SYMBOL (If applicable): NL	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER: AFOSR-88-0242
8c. ADDRESS (City, State and ZIP Code): Bld. 410, Bolling AFB DC 20332-5260	10. SOURCE OF FUNDING NOS: PROGRAM ELEMENT NO. 61102F; PROJECT NO. 2313; TASK NO. A7; WORK UNIT NO.
11. TITLE (Include Security Classification): Measuring Learning Ability by Dynamic Testing - Unclassified	
12. PERSONAL AUTHOR(S): Susan Embretson	
13a. TYPE OF REPORT: Final Report; 13b. TIME COVERED: FROM 88-8-1 TO 89-7-31; 14. DATE OF REPORT (Yr., Mo., Day): 89-9-30; 15. PAGE COUNT: 97	
16. SUPPLEMENTARY NOTATION	
17. COSATI CODES: FIELD / GROUP / SUB GR	18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number): Learning Ability; Dynamic Testing; Cognitive Testing

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

A criticism of traditional ability tests is that they are static, rather than dynamic, measures of intelligence. That is, they measure what the person has learned, but not neccessarily the capacity to learn. This project developed two tests of learning ability, spatial learning ability and mathematical learning ability, based on cognitive theory. In these tests, which consist of a pretest and two posttests, learning ability is the modifiability of a person's performance under conditions that change the cognitive load of the task, such as strategy training or cues. To solve some psychometric problems in measuring change (i.e., the inequivalencies of raw change at different initial performance levels and the unreliability of change scores), the multidimensional Rasch model for learning and change (Embretson, 1987; 1989a; 1989b) was used to estimate learning abilities. Further, the tests were counterbalanced for the stimulus features that influence processing difficulty to asure cognitive equivalency and to observe the impact of strategy training and cues on the mental models used in the tasks. (continued on back)

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT: UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION: Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL: John Tangney	22b. TELEPHONE NUMBER (Include Area Code): 202-767-4278; 22c. OFFICE SYMBOL: NL

DD FORM 1473, 83 APR — EDITION OF 1 JAN 73 IS OBSOLETE

Unclassified
SECURITY CLASSIFICATION OF THIS PAGE

89 12 04 127

Block 19 continued

Three goals were accomplished for each test: 1) large sample data was obtained to calibrate the tests by the multidimensional Rasch model for learning and change, 2) the construct validity of the learning ability measurements was examined and 3) the cognitive theory underlying the tasks in each test was extended. Although the results on mathematical learning ability were not particularly strong, the measurement of spatial learning ability was strongly supported.

Measuring Learning Abilities by Dynamic Testing Procedures

Susan E. Embretson[1]

University of Kansas

The purpose of this project was to provide psychometric data for two learning ability tests that are based on cognitive theory. Two complex psychometric tasks, spatial folding and mathematical problem solving, were presented under conditions that were designed to isolate specified aspects of cognitive processing or declarative knowledge. Both tasks have been studied intensively by the information processing approach so that a rich theoretical foundation existed to build cognitive models of the tasks, which specify the processes, strategies and knowledge structures that are involved in problem solving.

The major goal for this project was to provide a foundation for measuring learning ability for spatial reasoning and mathematical reasoning. Static tests of ability do not necessarily measure learning ability. Even at the first conference on intelligence testing, Dearborne (1921) noted that "...most tests in common use are not tests of the capacity to learn, but are tests of what has been learned." This same complaint has continued into modern intelligence research. Resnick and Neshes (1984, p. 276) point out that viewing static tests as measures of learning processes is based on the (probably faulty) assumption that "...the processes required for performance on the tests are also directly involved in learning." These concerns suggest that the optimal way to measure learning ability is to measure it directly, as a response to cues

Dist
A-1

or instruction.

In this project, learning ability was measured by assessing the modifiability of the person's performance on an ability test under conditions that are designed to increase performance levels by changing the cognitive load of the task. It should be noted that this definition differs in two important ways from the early studies on learning ability (e.g., Woodrow, 1938; 1940) that had negative findings. First, performance is measured on complex tasks (i.e , ability test items) with only a few replications. The early studies focused on rather simple tasks that were measured over many trials. When individual differences converge to an asymtotic level on the simple tasks, performance changes are highly correlated with initial performance levels, which accounts for Woodrow's (1938; 1940) findings. Second, the experimental conditions are designed to change specific cognitive processes, strategies or knowledge structures. The early studies did not have cognitive models of performance and therefore measured change that occurred spontaneously over undirected practice.

Several contemporary studies, using diverse ability tests, support the psychometric utility of measuring learning ability from performance modifiability (Budoff, 1987; Budoff & Hamilton, 1974; Carlson & Weidl, 1980; Campione, Brown, Ferrara, Jones & Steinberg, 1985; Embretson, 1987a; 1987b; Feurerstein, 1979, 1980). These studies use a dynamic testing procedure, which involves the administration of test-related training followed by a measurement of performance change. When successfully applied, dynamic testing procedures produce large increases in performance levels without adversely influencing test validity. In fact, some studies

have found that predictive validity is actually increased. Most directly relevant to this project, Embretson (1987a) found that performance modifiability on the spatial folding task had incremental validity for predicting vocational training in computer operations, when added to the initial ability.

These contemporary results on measuring learning ability oppose a traditional view of test developers; namely, that test- related training, called 'coaching', has a negative impact on predictive validity. Instead, the results support the view of the Russian learning theorist Vygotsky (1978). Vygotsky proposed the decomposition of a subject's performance level into two components: actual developmental level (a person's initial performance) and the zone of potential development (indicating the amount of learning that is within reach of a person's present performance). For the spatial aptitude task (Embretson, 1987), the locus of the increased predictive validity appeared to be situated in subjects' ability gains (i.e., zone of potential development), as separated from subjects' initial ability level (i.e., actual developmental level). Thus, the feasibility of measuring the zone of potential development, or a person's learning ability for a spatial aptitude task was demonstrated.

Psychometric Issues in Measuring Learning Ability

Appropriate psychometric models for measuring learning abilities have been difficult to develop. Attempts to measure learning in the context of classical test theory (e.g., Harris, 1963) led to many seemingly irresolvable conflicts. For example, Bereiter (1963) noted that raw gain scores 1) confound the reliability of change with the reliability of the tests (i.e., the higher the pretest correlates with the posttest, the more

unreliable the change scores), 2) are not necessarily measured on the same scale for different initial levels of ability, and 3) contain a spurious negative element when estimated by difference scores. As for the latter, Lord (1963) showed that the regression effect, under the condition of no real change, produces the spurious negative correlation. In summary, these various problems all suggest that , under classical test theory, the meaning of change is highly influenced by the multivariate distributions of the pretest and posttest measures.

Some of these seemingly irresolvable problems in measuring change under classical test theory can be resolved by conceptualizing change in the context of item response theory. In item response theory, abilities are measured as latent response potentials for performance, in the context of specific items with known difficulties. That is, the response to each item is a function of the difference between a person parameter (person ability) and an item parameter (item difficulty). Learning effects may be incorporated into the item response model parameters if it is possible to place linear constraints on the item parameters, as in the linear logistic latent trait model (LLTM, Fischer, 1973). Fischer (1972) proposed a linear logistic latent trait model with relaxed assumptions (LLRA) directly incorporates the trend effects. Similarly, Sapda and McGaw (1985) describe how LLTM may be used to test hypotheses about the nature of change. An advantage of these models is that the parameter estimates are specifically objective and do not depend on population distributions of pretest and posttest scores. Unfortunately, however, these latent trait models are not appropriate for measuring learning ability, as they contain parameters for overall learning effects, but do not include parameters to measure

individiual differences in learning.

Recently, Embretson (1987; 1989a; 1989b) has proposed two multidimensional item response models that solve several problems (i.e., Bereiter, 1963) in measuring change. The models measure initial ability and one or more learning abilities as latent response variables in a multidimensional model of performance. This psychometric model is especially crucial for measuring learning ability from performance modifiability, since it eliminates both the significant scaling artifact and a reliability paradox, mentioned above, that occurs when change is measured by raw performance scores (or standard scores). This project applies these multidimensional item response models to estimate modifiabilities for both spatial ability and mathematical reasoning ability.

To understand the model, it is necessary to distinguish theoretically between several kinds of ability. Performance is governed by the "effective ability" that corresponds to the conditions under which the ability measurement is obtained. The conditions include both the amount of the preceding testing and the extent of training for the measuring task. Initial ability, which is the standard measurement procedure, is defined as a person's latent response potential when a minimal set of instructions and practice precedes the test. Learning ability, in contrast, is the change in latent response potential that occurs when extended instruction and practice precede the test. If more extended instructions or cues are administered, as in a dynamic testing procedure, then the effective ability also includes a learning ability. Thus, learning ability is a separate ability, defined by different conditions of measurement. It is a change in

a person's effective ability, not a permanent change in initial ability. Embretson (1989b) presents examples which demonstrate the feasibility of estimating the learning ability parameters from the multidimensional model for learning and change.

Project Goals

The general goal of the project is to develop measures for both spatial learning ability and mathematical learning ability. These goals can be organized under three subgoals: 1) calibrating parameters for learning ability according to the multidimensional item response model for learning and change, 2) examining the construct validity of the learning ability measurements and 3) elaborating the cognitive theory that underlies the measuring tasks.

Large sample calibration of learning ability tests. The primary goal for the project was to provide large sample psychometric data for the measurement of both spatial learning ability and mathematical learning ability. For both types of abilities, initial ability and two learning abilities were measured. The learning ability measurements were defined by the experimental designs that present either instruction or structured cues that influence specific processes or knowledge structures. For both tests, an initial performance measure was obtained when the tasks were administered without instruction or cues. Then, instruction or cues were given with the task according to an experimental design. When appropriately targeted, the short interventions (i.e., 15 to 20 minutes) should increase performance levels substantially, but do so differentially across individuals.

Thus, each subject was tested under three conditions for each task: 1)

the pretest, in which the task was presented without special instructions, 2) the first posttest, in which the task was presented following strategy training, and 3) the second posttest, in which the task was presented following a second training condition. For each individual, the initial ability was estimated from the pretest, while the learning abilities were estimated from the performance increases from the pretest to the first posttest, and then from the first posttest to the second posttest.

The multidimensional item response model for learning and change (Embretson, 1987; 1989) was used to calibrate the items for difficulty and to estimate abilities and modifiabilities. The model requires that each item be observed under every condition, but that each item is presented under only one condition for any subject. Thus, three groups were required to observe an item under every condition. To control bias, items and conditions of observation were counterbalanced (pretest, first posttest, second posttest). A Latin square design (see below) was used to counterbalance items and conditions over the three groups of subjects. Further, a small set of items was administered under the same condition (i.e., the pretest) to link item parameter estimates across groups.

Extend construct validity. The second goal was to extend the construct validity of the learning ability measurements by examining both construct representation and nomothetic span (see Embretson, 1983). Construct representation concerns the cognitive processes, strategies and knowledge structures that are involved in performance. Construct representation in this project was studied by mathematically modeling item accuracy and response time from stimulus properties that reflect the cognitive load of the item, according to a theoretical model for the task.

The theoretical models were developed on the basis of prior research and theory, and on empirical findings with the ability test versions of the spatial ability and mathematical reasoning tasks.

Because it was hypothesized that the experimental interventions not only increase performance levels, but also change the cognitive load of the task, having an empirically plausible theoretical model is very important. If the cognitive load of the task changes following intervention, then the weights of the various cognitive variables that operationalize the theoretical model should be significantly different between the pretest and posttests. Adequately testing this hypothesis requires that the stimulus features that influence the difficulty of cognitive processing be carefully balanced between the tests. Thus, the theoretical model guides the design of the test measures as well as defines a testable mathematical model of performance.

Nomothetic span, the meaning of the ability scores, was examined by both the internal properties of the tests (i.e., internal consistency, goodness of fit to an item response model) and by the relationship of the several ability scores to each other and to other measurements of individual differences. Specifically, the covariances of the performance levels over conditions of both spatial ability and mathematical reasoning were hypothesized to fit a model for a quasi-Weiner simplex (see Joreskog, 1979). Furthermore, an external validation measure was also available to examine the incremental validity of the modifiability measures over the initial ability. The external criterion measure was success in short intelligent tutoring system on electronics trouble-shooting (i.e., the Logic-Gates task).

Extend cognitive theory. The third goal was to extend cognitive theory on both complex spatial processing and mathematical reasoning by developing processing models for each task. Although theoretical foundations for both spatial processing and mathematical problem solving tasks exist, further work clearly is needed. The existing theories are far from complete. Furthermore, it is not clear to what extent the various theories are empirically plausible for the variants of the tasks that actually appear on ability, or that can achieve adequate psychometric properties.

Thus, this project began with studies to further develop the cognitive models of both tasks. These models further general knowledge about the tasks, and, more directly relevant, are needed to elaborate the construct validity of the learning ability measurements with respect to cognitive load. This implies that the cognitive models of task response time and accuracy will change after the intervention. An empirically plausible cognitive model for the task thus is needed to explicate the impact of the intervention.

Status of the Project

The primary goal of the project was to obtain a rich, large scale data base for measuring learning ability. Because the availability of large sample data from a computerized testing laboratory is a rare and valuable opportunity, the greatest effort in the project was placed into the careful design of the test measurements and the interventions prior to collecting data at Lacklund Air Force base. Thus, an intensive reanalysis of a prior study, three new studies and two pilot studies at Lacklund were completed before the actual testing in May and June at Lacklund Air Force Base. Many

additional analyses of the Lacklund data are planned. Most of the studies will be completed after the due date for this report. These plans will be included in this final report, in addition to the many results that already have been obtained.

Organization of the Report

The spatial learning ability and mathematical learning ability projects will be described separately in the report, as their theoretical bases are quite different. For each task, the following areas will be covered: 1) a summary of relevant theory and research will be given, including supporting research from the principal investigator's laboratory, 2) cognitive models for the task will be postulated, 3) studies that tested the cognitive model in this project will be presented, 4) the calibration data from the main Lacklund study will be presented and 5) studies relevant to construct validity will be presented. The section on the calibration data includes a detailed description of the cognitive design of the tests, results from the psychometric model fitting and norm development, and results on the construct validity of the learning ability measurements.

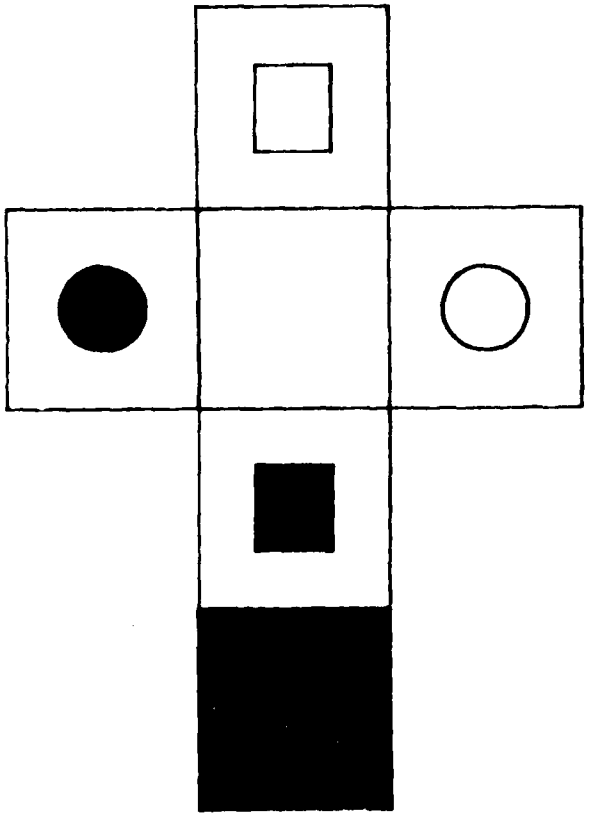
Measuring Spatial Learning Ability

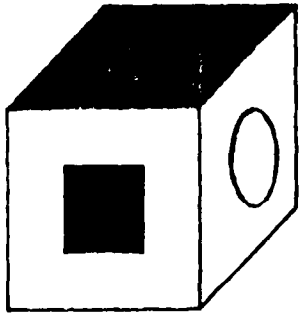
Cognitive Theory and Spatial Folding

Spatial tasks have well-established psychometric importance in measuring ability due to the differential validity of spatial ability to predict success in vocational classes such as graphics and engineering courses, shop courses and microcomputer operations (Egan & Gomez, 1982). Tasks that involve rotating figures in two or three dimensions have a long history in intelligence tests. Early tests of spatial ability ranged from very simple two dimensional tasks in Thurstone and Thurstone's (1941) studies on the primary mental abilities to more complex tasks that involved several stages, such as the Minnesota Paper Form Board. A major contemporary test, the Differential Aptitude Test (DAT), uses a complex task, three-dimensional spatial folding, to measure spatial ability.

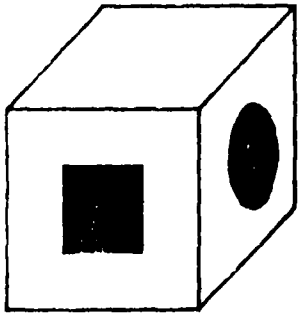
Several tasks that appear on spatial ability tests have been studied in the laboratory. A major finding is that the angle of rotation is linearly related to response time for comparing two objects. Studies on both two- dimensional rotation (Cooper & Shepard, 1973) and three-dimensional rotation (Shepard & Metzler, 1971) as well as more complex tasks, such as comparing cube (Just & Carpenter, 1985) and figure construction (Pellegrino, Mumaw & Shute, 1985) yield rather similar results. First, the angle of rotation is linearly related to response time for comparing two objects. Second, although the exact cognitive models differ between various tasks that have been studied, the processes of anchoring, rotating and confirming are typically contained in the models.

The spatial folding task, such as appears on the DAT, has not often been studied in the laboratory. Figure 1-1 shows a sample item. Only one

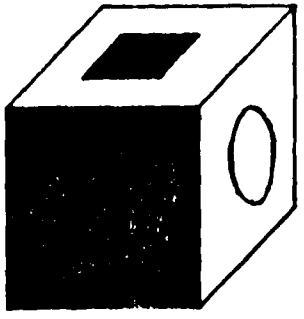




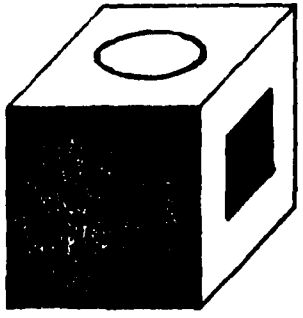
1)



2)



3)



4)

alternative shows a three-dimensional view that would result from folding the sides of the stem downward to construct the figure. An early study (Shepard & Feng, 1972) suggested that the number of surfaces carried influences processing response time. This is not surprising, when compared to the findings from two-dimensional rotation tasks, as folding can be considered a depth rotation of the surfaces to construct the object.

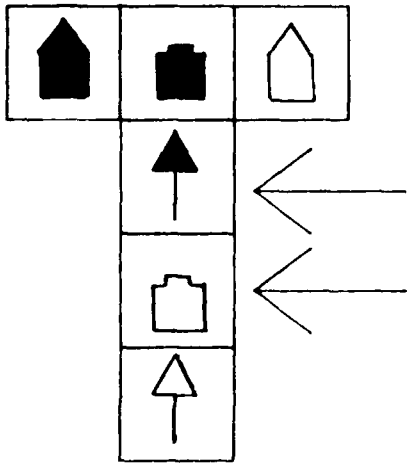
The next section proposes a model of spatial folding, the attached folding model, that is consistent with Shepard and Metzler's (1971) results as well as with the general processing models from other spatial tasks. Two alternative models will be proposed as well. One model is a direct folding model, which contains a cube comparison process (e.g., Just and Carpenter's (1985) theory) after folding the stem. The other model is a verbal-analytic model for processing the spatial folding task, which is based on some simple propositions of the relative locations of the marked sides.

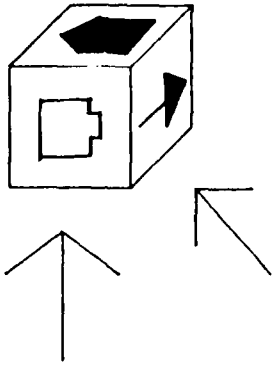
Models for Spatial Folding

The attached folding model. In the attached folding model it is postulated that the task is solved by a spatial strategy, but that the strategy minimizes the load on spatial processing resources by folding the stem to fit a particular alternative. It is postulated that a person mentally attaches two adjacent sides of the unfolded stem to the alternative, and then folds until the third surface appears in place.

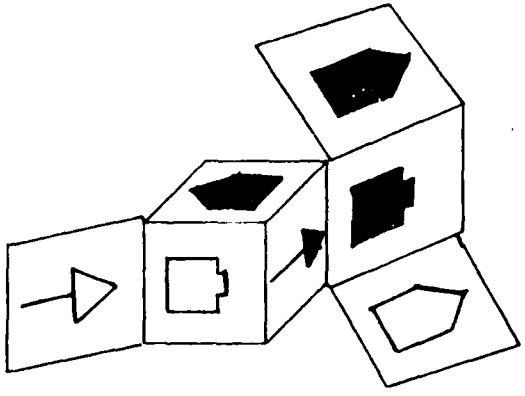
Figure 1-2 shows the processes of anchoring, rotating and folding in the attached folding model. In Figure 1-2, the search for an anchoring point entails encoding the folded alternative, and then locating an adjacent pair of markings on the unfolded stem that appears on the

SELECT ANCHORS

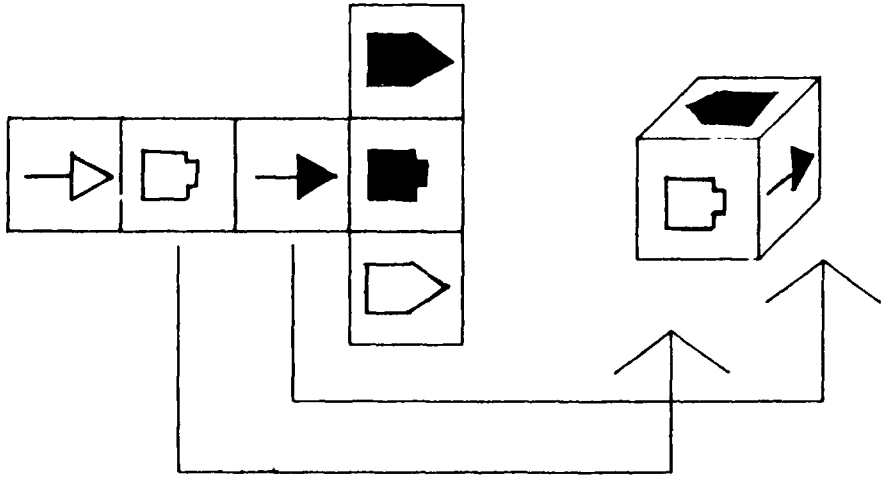




FOLD



ROTATE AND ATTACH



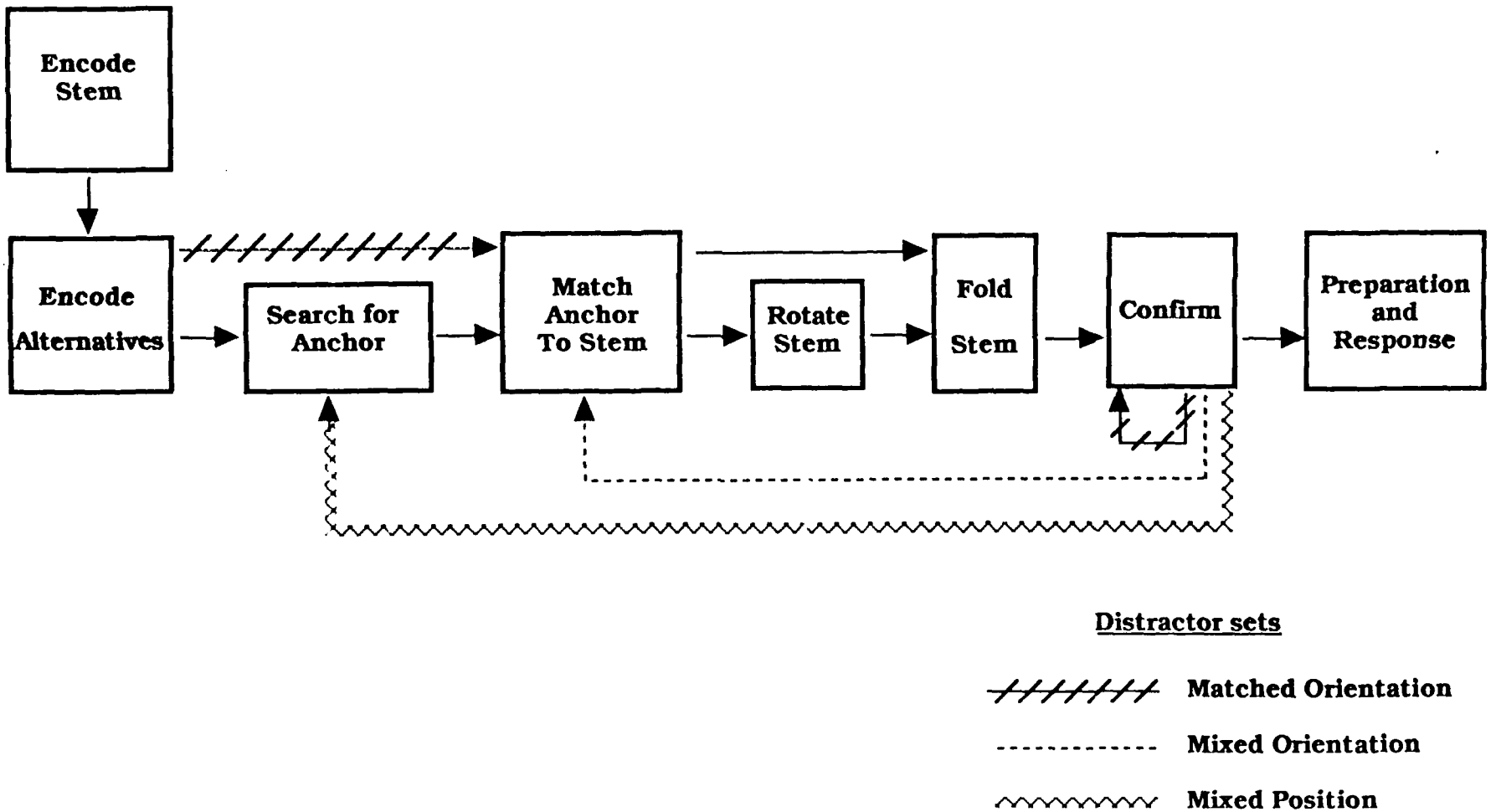
alternative. Then, the adjacent sides on the stem are attached to the folded alternative. Attaching may involve rotation. Figure 1-2 shows a mental rotation of 90 degrees to attach the stem to the folded alternative. Then, the third side is folded. A final major process (not shown) is confirming that the same markings appear on the folded stem and the folded alternative.

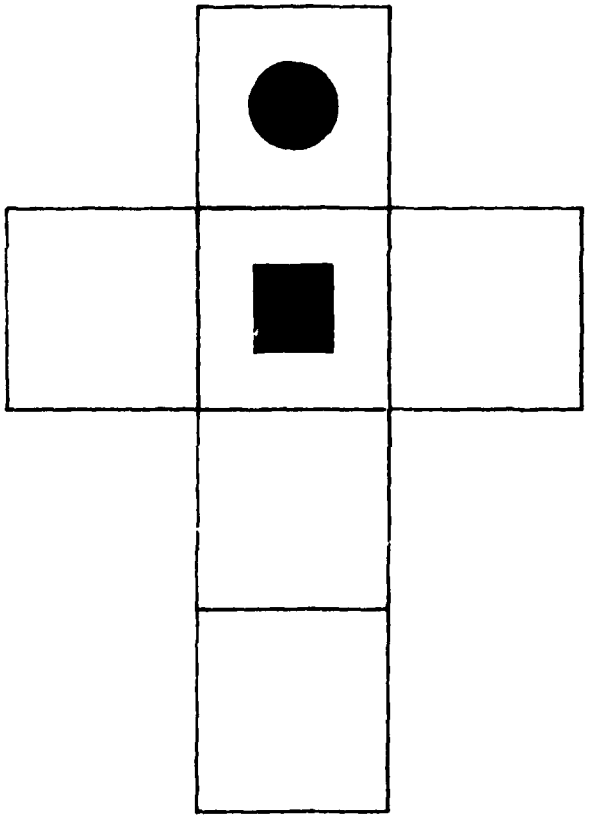
Figure 1-3 presents a schematic of information processing with the attached folding model. The model contains both invariable and variable processes, depending on the distractor set. Encoding is invariable, with separate encoding of the stem and alternative(s). Then, a variable process in which an alternative is searched for an adjacent pair is undertaken. If all the distractors match the correct answer except for the third side, an extended search for the adjacent pair is not necessary. In this case, the same adjacent pair always appears in the same position on every alternative.

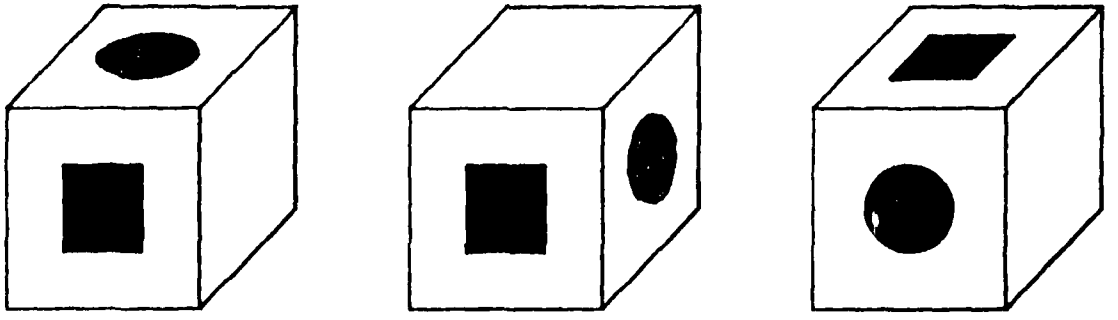
Next, the adjacent pair is located on the unfolded stem. The unfolded stem is then mentally attached to the folded alternative. Rotation may be required to attach the stem, depending on the orientation of the markings on the alternative. Figure 1-2 shows a 90 degree rotation. Finally, the third side is mentally folded into place, and then the markings are confirmed on the stem with respect to the folded alternative.

The next stages depend on the nature of the distractor set. If the task is presented in the verification format, with only one alternative, then the preparation and response stage is entered, with the response of "true" given if the confirmation was successful; "false" if confirmation failed. If several alternatives are given, the model specifies that they

Attached Folding Model







0°

90°

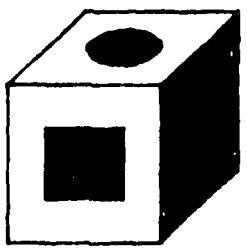
180°

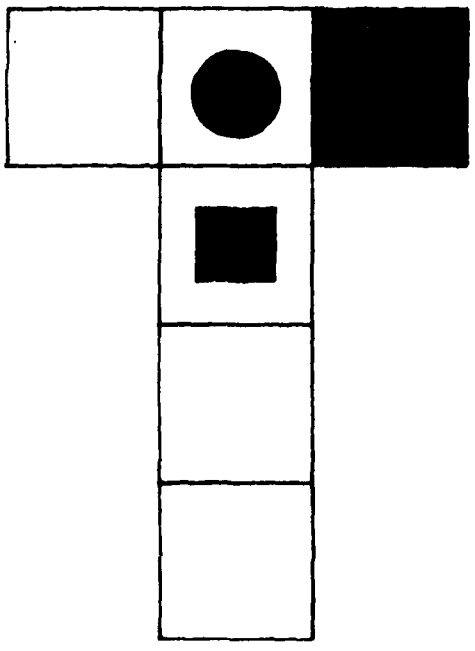
are processed exhaustively. If all alternatives vary only in the third folded side, no additional rotating or folding of the stem is attempted. Instead, the third side of the mentally folded stem is held in memory and then is compared to each alternative. The alternative that is not disconfirmed is then selected as the correct answer. Thus, confirmation will continue exhaustively.

If the alternatives contain the same adjacent pair, but in rotated orientations, additional matching to the stem and attaching, rotating and folding are undertaken for each alternative. This continues exhaustively until all alternatives are processed. The alternative that is not disconfirmed is selected as the correct answer. If the alternatives vary in both orientation and the markings on the adjacent pair, each alternative must be fully processed. Thus, processing begins for each new alternative by searching for an anchor and continuing to confirmation.

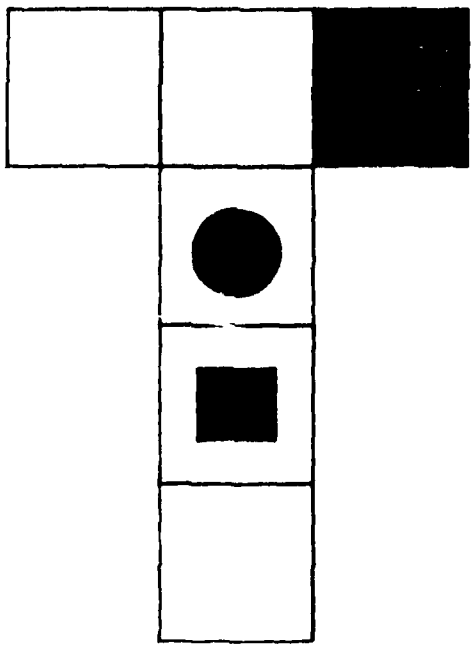
The encoding process is probably manipulated by the complexity of the markings on the sides of the cube. In the attached folding model, it is postulated that encoding involves images, rather than proposition. The confirmation process probably is also manipulated by the complexity of the markings. In Figure 1-2, all markings appear different from different sides, so that confirmation is more difficult. However, if the markings appear the same, as all the markings in the unfolded cube on Figure 1-1, confirmation is easier.

Attaching and folding involve mental rotations, so the manipulation of their difficulty will be described more completely here. Figure 1-4 shows how the difficulty of attaching can be varied by the degrees of rotation of the stem to the alternative. Figure 1-5 shows an unfolded

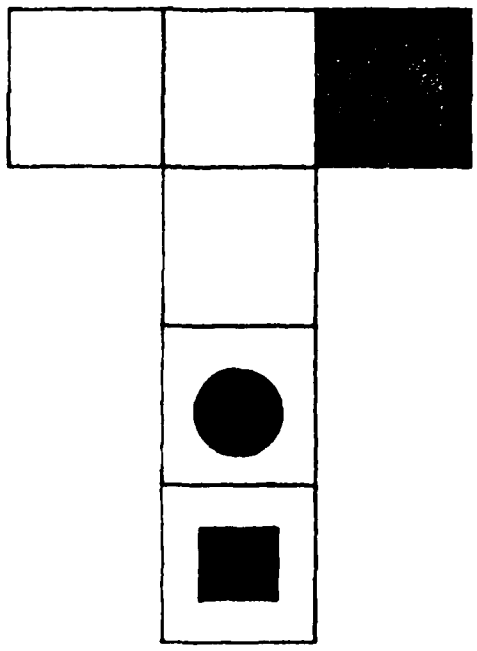




1 Surface



2 Surfaces



3 Surfaces

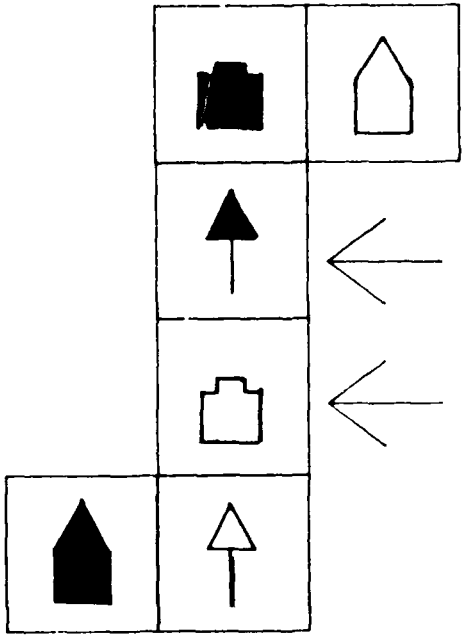
stem and three different correct alternatives that vary in attachment difficulty. The stem can be directly overlayed on the first alternative without any rotation. The second and third alternatives, in contrast, require a 90 degree and 180 degree rotation of the stem, respectively.

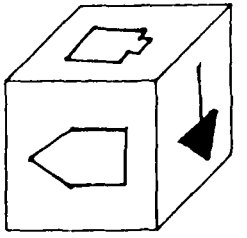
Figure 1-5 shows how the difficulty of folding can be varied. Shown in Figure 1-5 are one correct alternative and three different stems. The first stem is the easiest because, if the shaded circle and shaded square are attached, only one surface needs to be carried to complete the folding. However, if the sides that contain these two markers are attached for the second stem, two surfaces must be carried mentally to bring the third side into place. The last stem requires that three surfaces be carried.

The direct folding model. The direct folding model is postulated to be more demanding on spatial processing resources. In the direct folding model, it is postulated that a person folds the stem fully before examining the alternatives. Then, the person enters a cube comparison process to compare the mentally folded figure to the alternative. It is postulated that the comparison process, although it involves a mentally folded and an actually folded figure, involves processing as described in Just and Carpenter's (1985) theory of cube comparison. This strategy is assumed to demand greater spatial processing resources because the folded stem must be held in memory while rotating to fit the alternative.

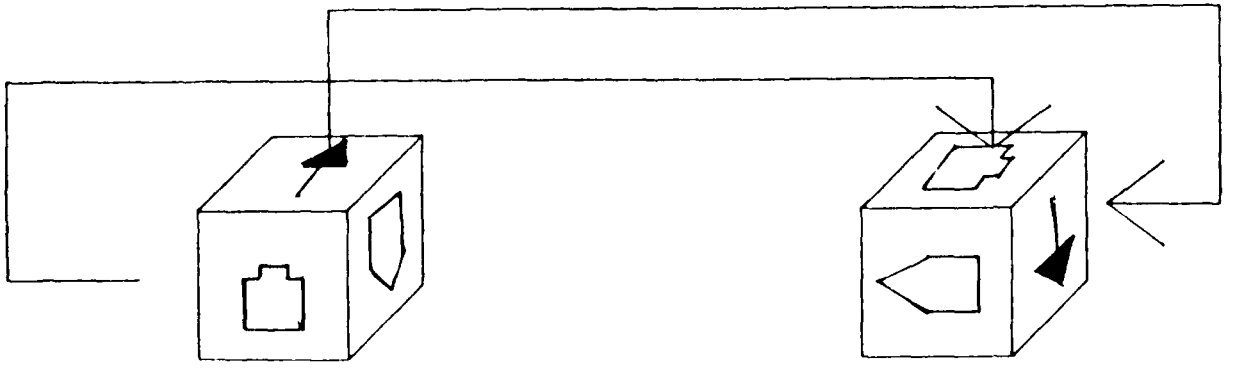
Figure 1-6 shows the major processing stages for the direct folding model. It is assumed that the stem is folded along the most balanced point, such that the edge of greatest balance lies between the front and the top of the mentally folded cube. Figure 1-6 shows the two sides that surround the folding edge for the example stem. After the stem is folded,

SELECT AND FOLD





ROTATE



then the person searches for an anchor and then rotates the mentally folded figure to the same orientation as the alternative.

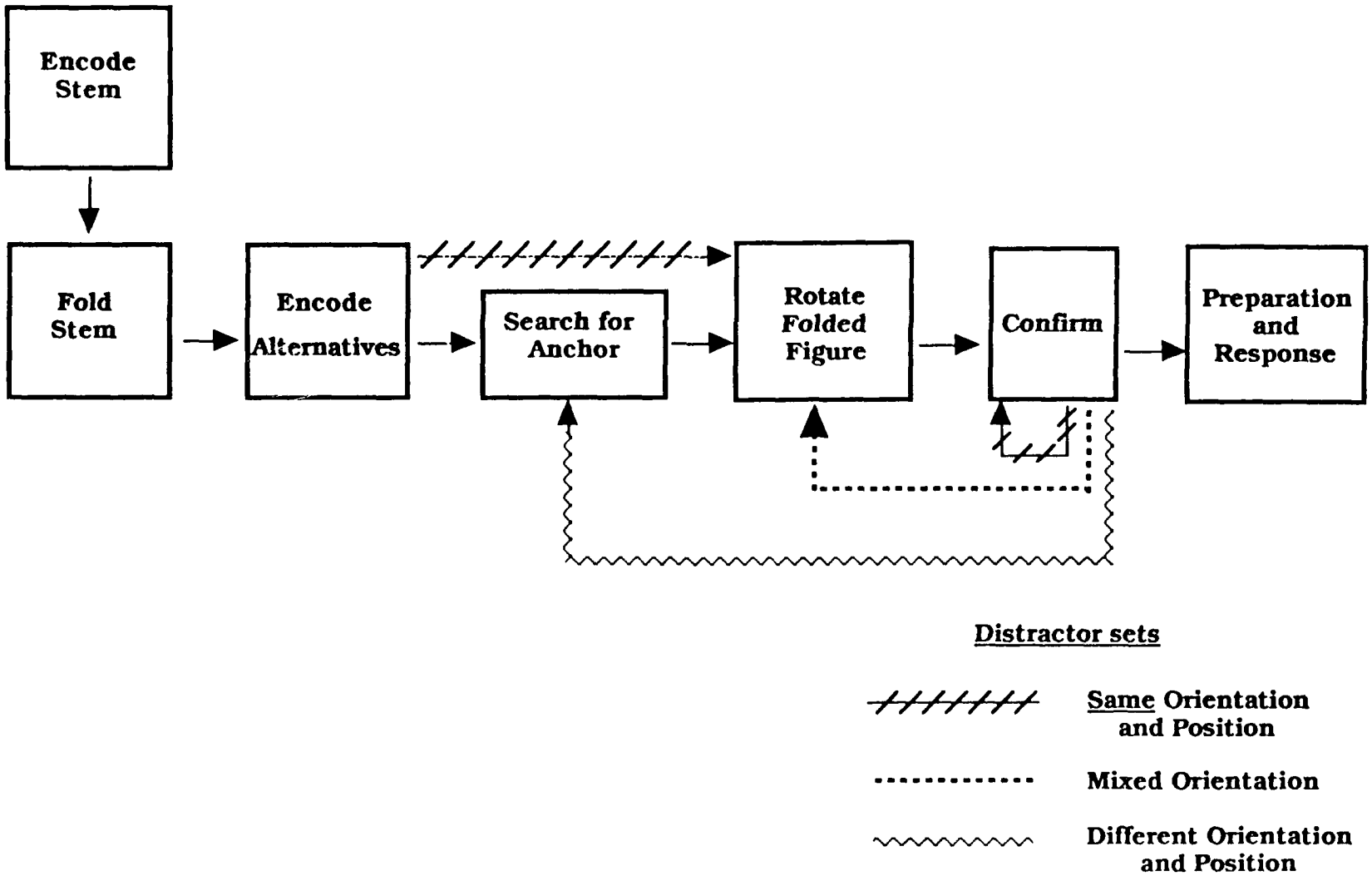
Figure 1-7 presents a schematic representation of the direct folding model. It can be seen that the distractor set influences the processing stages entered as in the attached folding model. However, since the stem is folded without respect to the alternatives, the successive stages involve only rotation of the mentally folded figure.

The verbal analytic model. In the verbal model, it is postulated that the unfolded figure is compared to a folded alternative by elaborating propositions about the relative locations of the markers. Such propositions may have a very elementary form, such as the propositions about location in the Chase and Clark (1972) task. For example, on Figure 1-8, two elementary propositions about the markings on the first alternative are: 1) "The white arrow is above the black rocket", and 2) "The black rocket is beside the black anchor". These same relationships hold on the stem, so no conversion is needed.

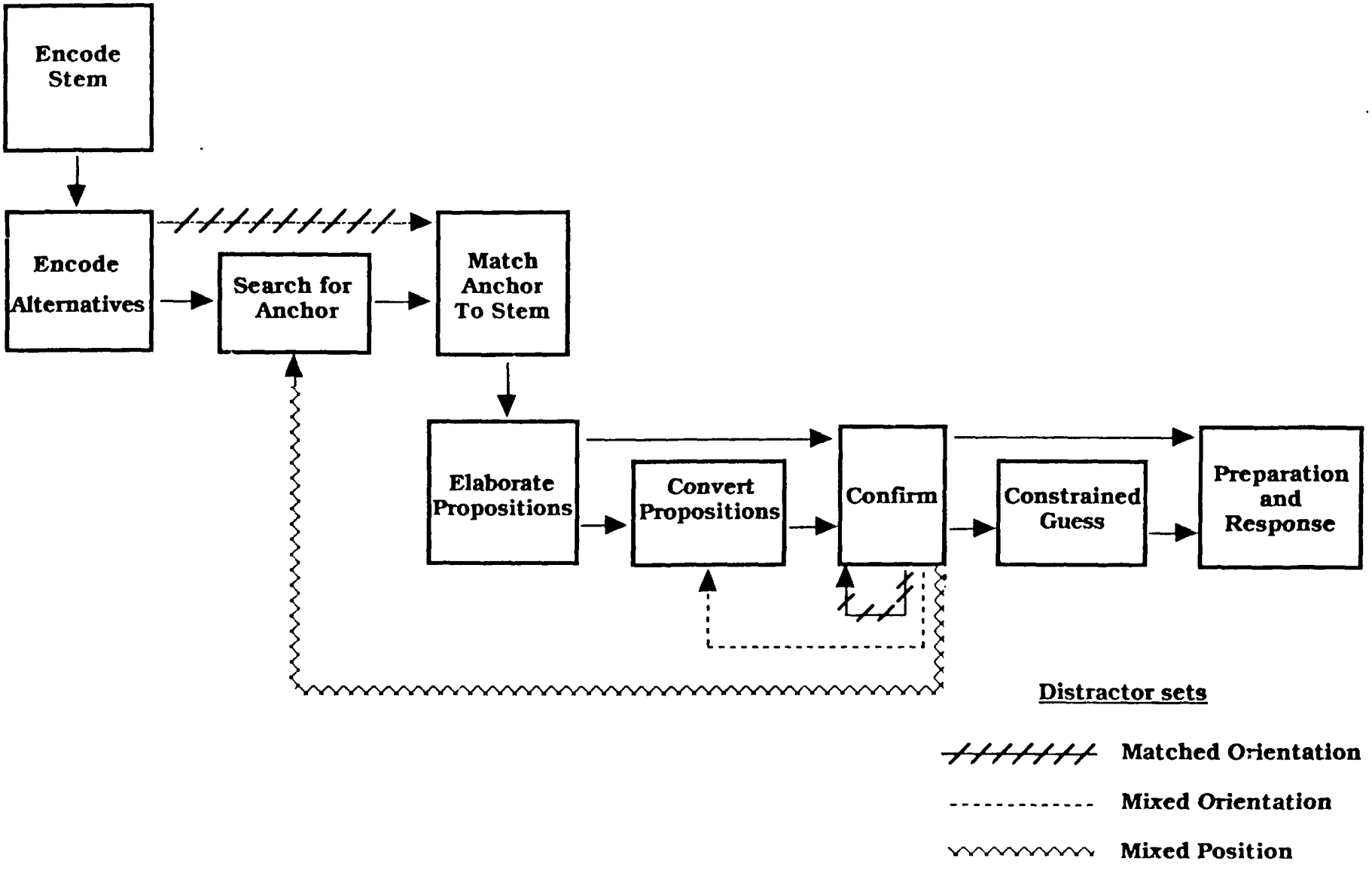
For the second alternative, the propositions must be converted, as the sides have been rotated 180 degrees. Now, the relationship of the white arrow and the black rocket is reversed (i.e., the white arrow is below the black rocket) but the black rocket is still beside the black anchor. Notice, however, that this relationship is not sufficient to confirm the alternative. The third alternative shows these relationships, but is not correct. So, a positional relationship, "The black rocket is to the right of the black anchor", must be added.

A position relation may also lead to confirming the first alternative on Figure 1-8. Here, the black rocket is below the white arrow, and the

Direct Folding Model



Verbal Model



black rocket is to the right of the black anchor, as on the stem. Again, this is not a sufficient relationship, however, because the fourth alternative has the same properties but cannot be made by folding the stem.

Thus, the verbal analytic model contains a constrained guessing stage in the case where the propositions are not sufficient. Figure 1-9 shows a schematic representation for the various stages in the verbal model.

Previous Studies

Modeling the DAT Spatial Folding Task

Embretson (1987) modeled item difficulty on the DAT, in the context of a study to examine the measurement of learning ability. In the study, the linear logistic latent trait model (LLTM) was used to model item responses. LLTM is a psychometric model of item responses that permits a mathematical model of the task to replace estimates of item difficulty. The data set included a pretest, an intervention and a posttest. Although the mathematical model showed changes from pretest to posttest, indicating that the cognitive basis of performance had changed, the model did not clearly operationalize a theory of processing.

The data from Embretson (1987) were reanalyzed to examine the empirical plausibility of the attached folding model for the spatial folding task as it appears on a spatial ability test. Item response data were available at both pretest and posttest for 45 DAT spatial folding items. The items were scored according to the attached folding model. First, the degrees of rotation for attachment and the number of surfaces carried in folding the stem to the correct alternative were scored. Second, the degrees of rotation and the number of surfaces carried to falsify the distractors was also scored. The maximum degrees of rotation

and number of surfaces carried for a distractor were retained, as were the number of distractors that could not be anchored to the stem. The latter case includes items in which one or more markings on a distractor do not match the markings on the stem or in which adjacent markings on the distractor are not adjacent on the stem. Also, the difficulty of the confirmation stage was scored as the existence of directed markings (i.e., they appear differently when viewed from different directions).

Table 2-1 shows the LLTM weights for the variables of the attached folding model. These weights are comparable to unstandardized regression coefficients for modeling a log odds (ln(P/1-P)) scale of item difficulty. The LLTM weights for unanchorable distractors, degrees rotation to the hardest distractor, number of surfaces carried to the worst distractor, and directional markings all had significant unique contributions to predicting item difficulty in the pretest. Thus, items are difficult when they contain no unanchorable distractors, require extensive rotation and several surfaces to be carried to falsify a distractor and contain directed markings. The level of prediction was comparable to a multiple correlation coefficient of .67. For the posttest, the weights for unanchorable distractors changed significantly. Furthermore, other weights showed marginal changes from pretest to posttest. The level of prediction also decreased, which was comparable to a multiple correlation coefficient of .53.

These results suggest that the attached folding model provides at least moderate prediction of item difficulty for the DAT spatial folding items. High levels of prediction, although highly desirable, should not be expected, however. The spatial items in the DAT were not constructed to

Table 2-1

Impact of Instruction on Cognitive Processing

of Spatial Ability Items

	Pretest			Posttest			
Variable	η	SEη	r	η	SEη	r	tdiff
Anchoring							
Unanchorable Distractor	-.29**	.05	-.27	-.06	.06	-.06	2.94
Degrees Rotation to Key	-.09	.08	.22	-.24**	.09	.07	1.25
Degrees Rotation to Distractor or Stage (Maximum)	.68**	.06	.49	.66**	.07	.49**	.24
Folding							
Surfaces Folded to Key	-.01	.04	.08	-.04	.04	.06	.54
Surfaces Folded to Distractor or Stage (Maximum)	.16**	.04	.34	.13**	.05	.21	.47
Confirming							
Directional Markers	.33**	.10	.32	.09	.12	.19	1.78

represent variants in item difficulty as defined by a theory. The variables of the attached folding model, therefore, are not explicitly counterbalanced in the DAT, so their effects cannot really be untangled. An inspection of the intercorrelations of the variables indicated that, in fact, the major sources of processing difficulty in the attached folding model, the degrees of rotation and number of surfaces carried, were not independent in the items.

Furthermore, the multiple distractors in the DAT are apparently developed according to mixed criteria. The items vary from containing very easily falsifiable distractors (e.g., unanchorable to the stem) to containing distractors that require extensive processing to falsify. Finally, for many items, the markings on the sides are redundant. Thus, there may be more than one way to attach the stem to the folded alternative and therefore their demand on spatial processing may vary. So, different processing strategies may exist for the same item, which will create instability in the processes that are applied. Although, in theory, this variability could be embedded in a mathematical model, it creates problems in testing the more simple propositions of the attached folding model.

Thus, considering the various complexities in the DAT items, the magnitude of prediction achieved indicates that the attached folding model is empirically plausible for the spatial folding task. A more convincing test of the model, however, requires more carefully constructed tasks.

Modeling the Spatial Folding Task with Unique Sides

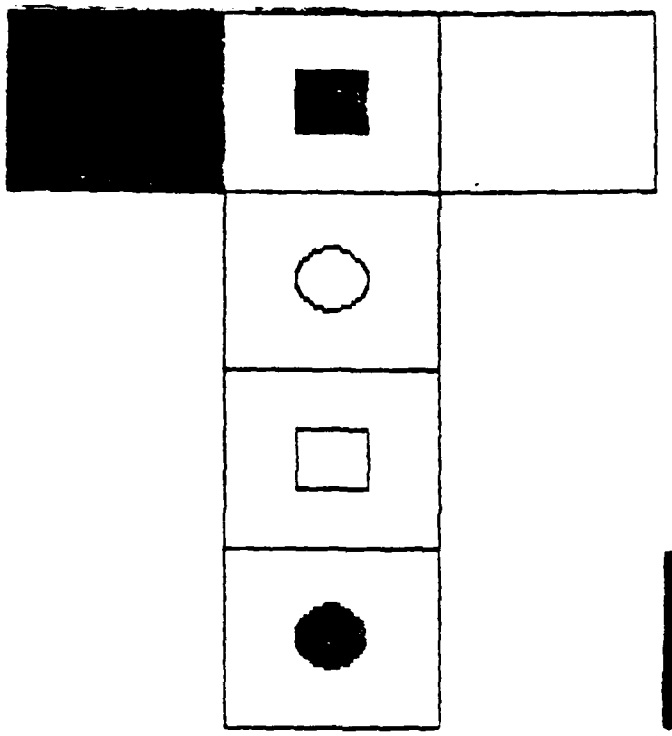
Embretson and Skube (1988) constructed items to provide a less confounded test of the attached folding model for the spatial folding task.

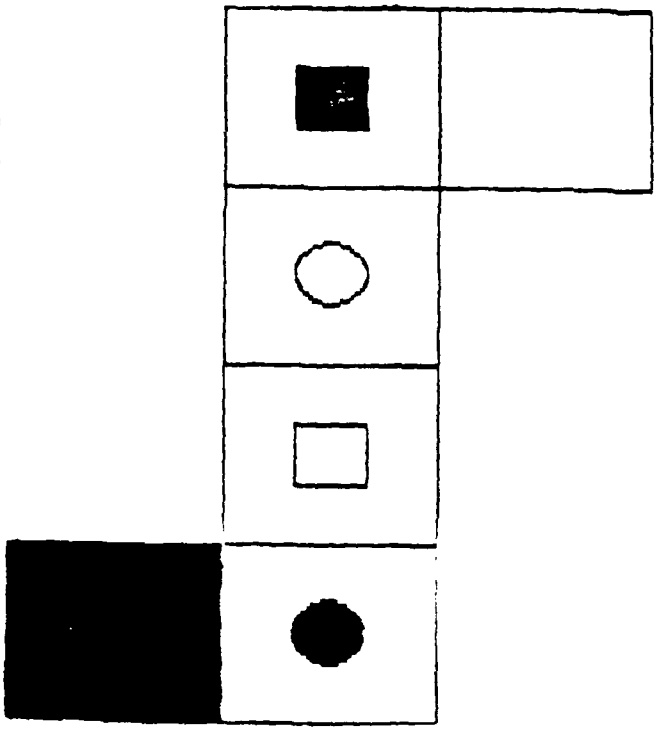
Several aspects of the task were simplified. First, only cubes were studied. Second, a verification format was studied, rather than the multiple choice format of the DAT. Thus, the process of confirming the correct answer could be examined without the confounding influence of the distractor context. Third, unique markers were placed on each side. This reduces the impact of multiple strategies for attaching the stem to the folded alternative.

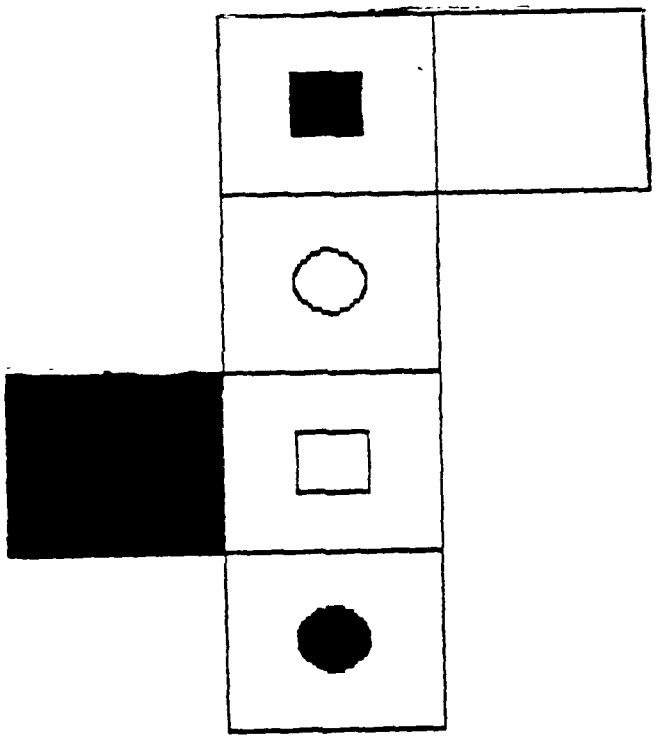
Figure 3-1 shows the four shapes for the unfolded stem that were used in the study. The main interest was in the true problems, but about one-third of the problems were false to prevent response bias. For true problems, the degrees of rotation (0, 90 or 180 degrees) was crossed with the number of surfaces carried (1,2 or 3), using four shapes and two stimulus sets (with different configurations of markings). Thus, for true problems, 72 items were generated according to a Degrees(3) X Surfaces(3) design with eight replications.

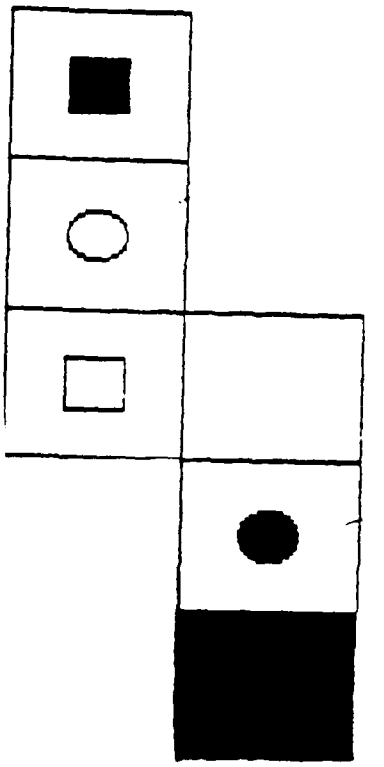
For false problems, the Degrees(3) X Surfaces(3) design generated 36 problems. In the false problems, the stimulus sets were counterbalanced rather than crossed with the problem variants. Additionally , three false problems with unanchorable sides were created for each shape, thus yielding 12 additional items.

The 120 items were administered in a random order to 31 college undergraduates who participated in the experiment to fulfill a course requirement. Within subject analyses of variance were performed on response times by taking the geometric means over stimulus set and shape. Both total response time and correct response were analyzed. Within subject analysis of variance were performed on accuracy by calculating







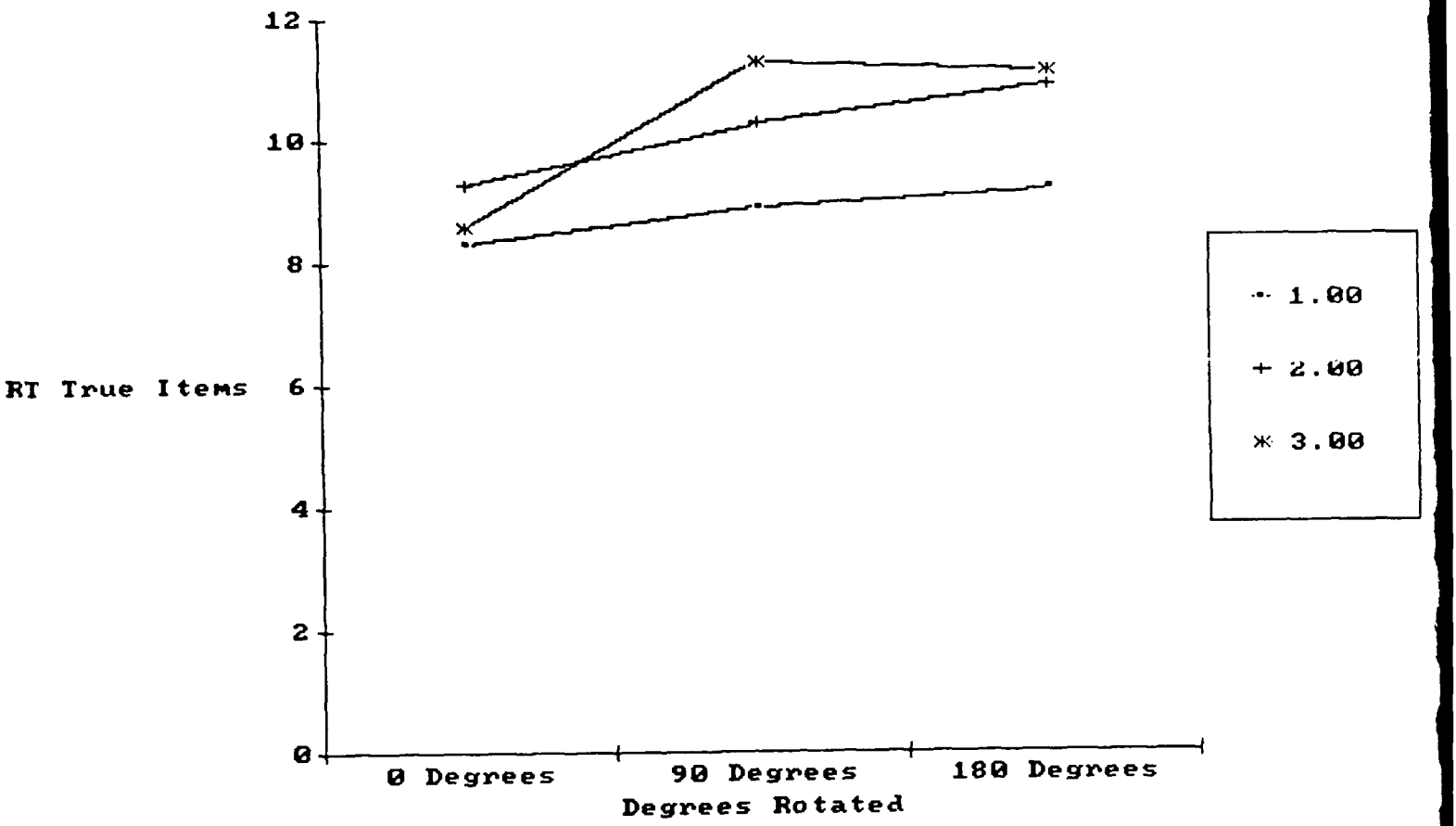


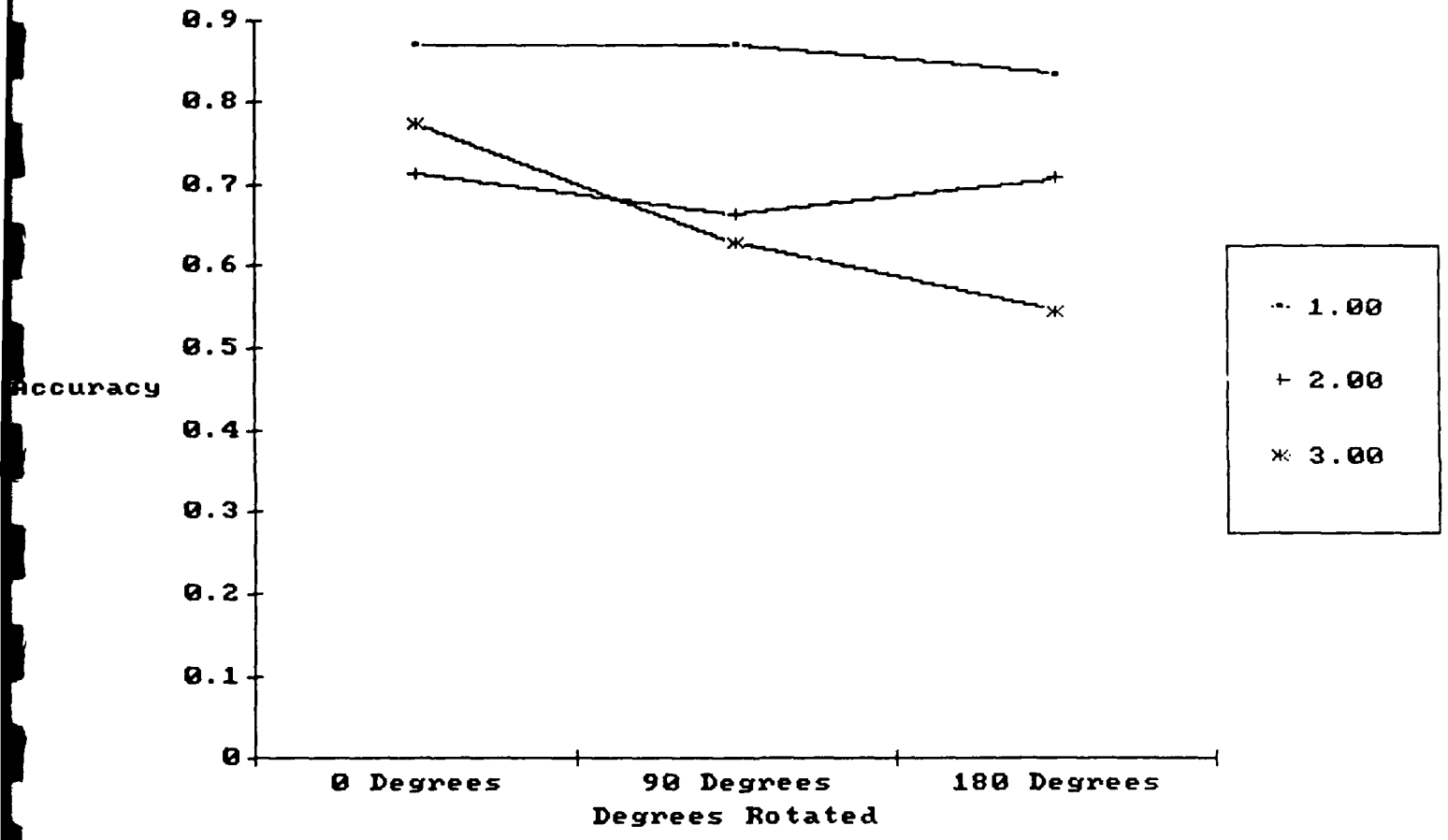
proportion correct over stimulus set and shapes.

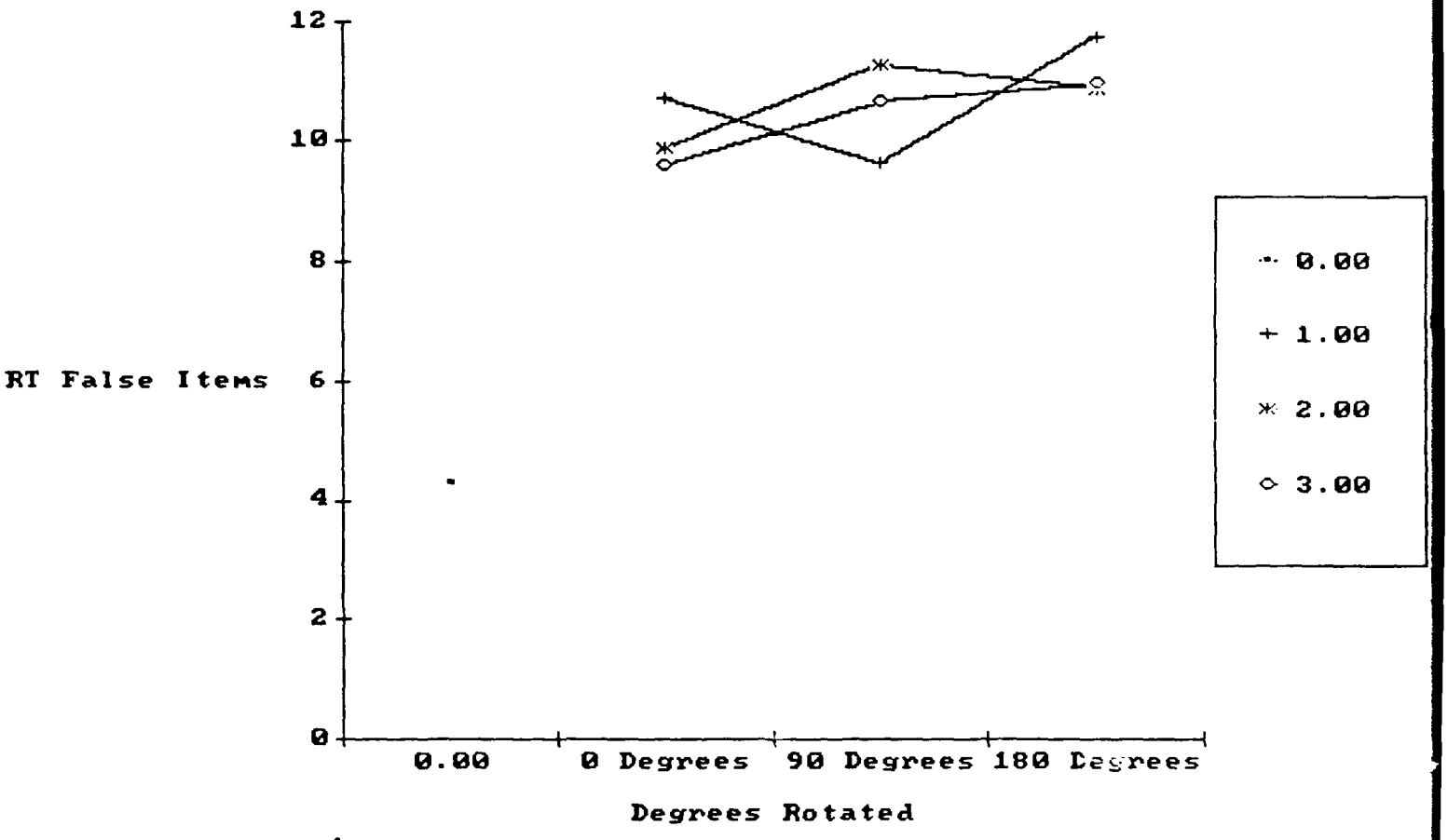
Figure 3-2 shows the mean total response times for true items, by condition. It can be seen that response time generally increases with the degrees of rotation ($p<.001$) and the number of surfaces carried ($p<.001$). However, the interaction of degrees of rotation and number of surfaces carried was significant ($p<.001$). Figure 3-2 indicates that the main contribution to the interaction disturbance is the condition of three surfaces carried. In this condition, which is generally harder, problems are relatively easy at zero degrees of rotation. An analysis of correct response times yielded the same pattern of significant effects and the same pattern of means.

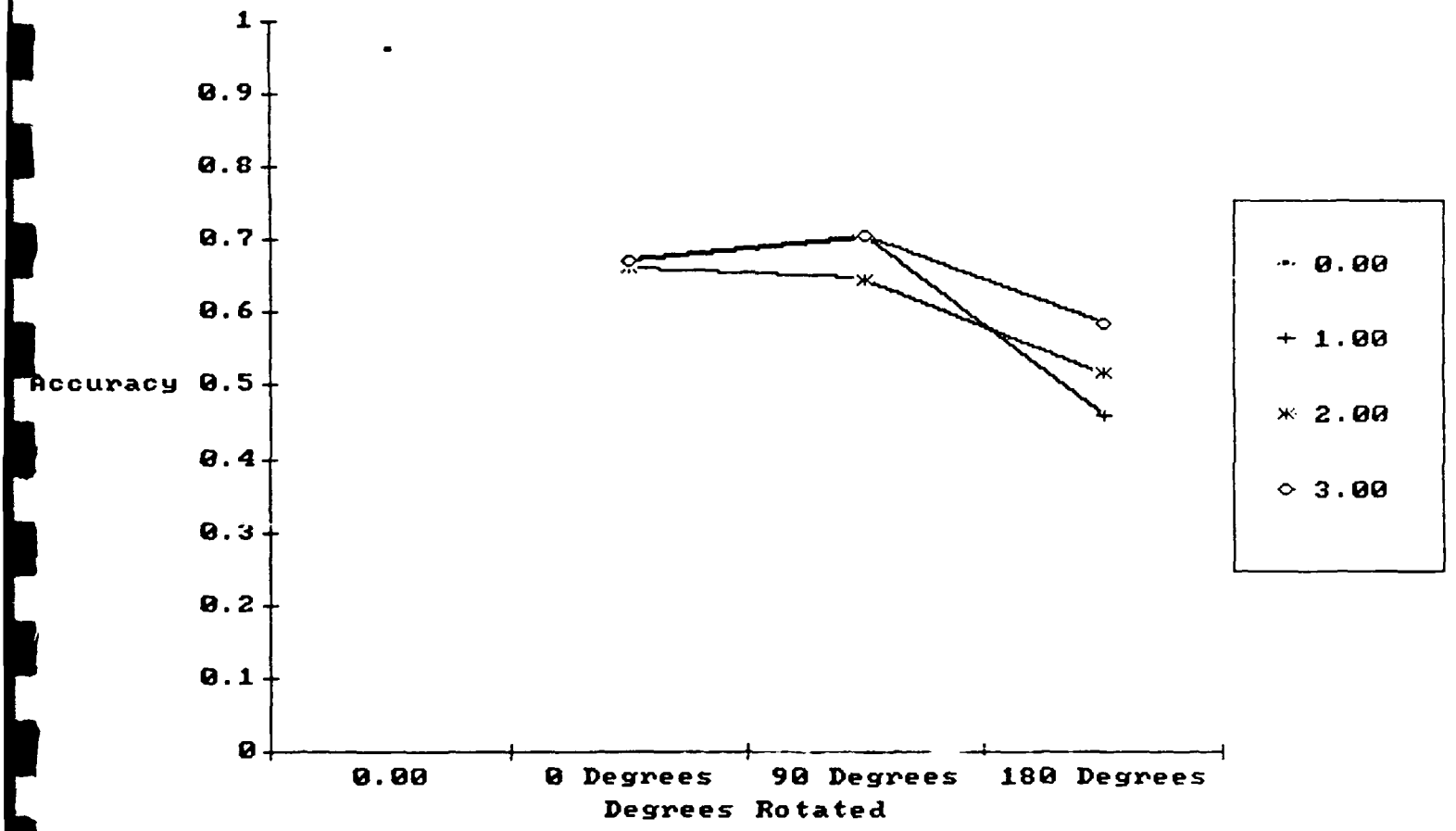
Figure 3-3 shows the mean accuracy by condition. Significant main effects were observed for both the number of surfaces carried ($p<.001$) and the degrees of rotation ($p<.001$). However, as for response time, the interaction of degrees of rotation with surfaces carried was also highly significant ($p<.001$). Figure 3-3 shows that the three surface problem was relatively easy at zero degrees of rotation.

The false problems were also analyzed. Figure 3-4 shows the mean total response times. A significant effect was observed for degrees of rotation ($p<.001$), but not for surfaces carried ($p>.05$). However, the interaction of surfaces and degrees of rotation was significant ($p<.01$). The one surface problems show a different pattern of relationship with degrees of rotation than the two or three surface problems. Also shown in Figure 3-4 is the response time for the unanchorable distractors. It can be seen that these problems were solved very rapidly in comparison with the other problems.


12
10
8
6
4
2
0
RT True Items
0 Degrees
90 Degrees
180 Degrees
Degrees Rotated
1.00
2.00
3.00


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Accuracy
0 Degrees
90 Degrees
180 Degrees
Degrees Rotated
1.00
2.00
3.00


12
10
8
6
4
2
0
RT False Items
0.00
0 Degrees
90 Degrees
180 Degrees
Degrees Rotated
0.00
1.00
2.00
3.00


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Accuracy
0.00
0 Degrees
90 Degrees
180 Degrees
Degrees Rotated
0.00
1.00
2.00
3.00

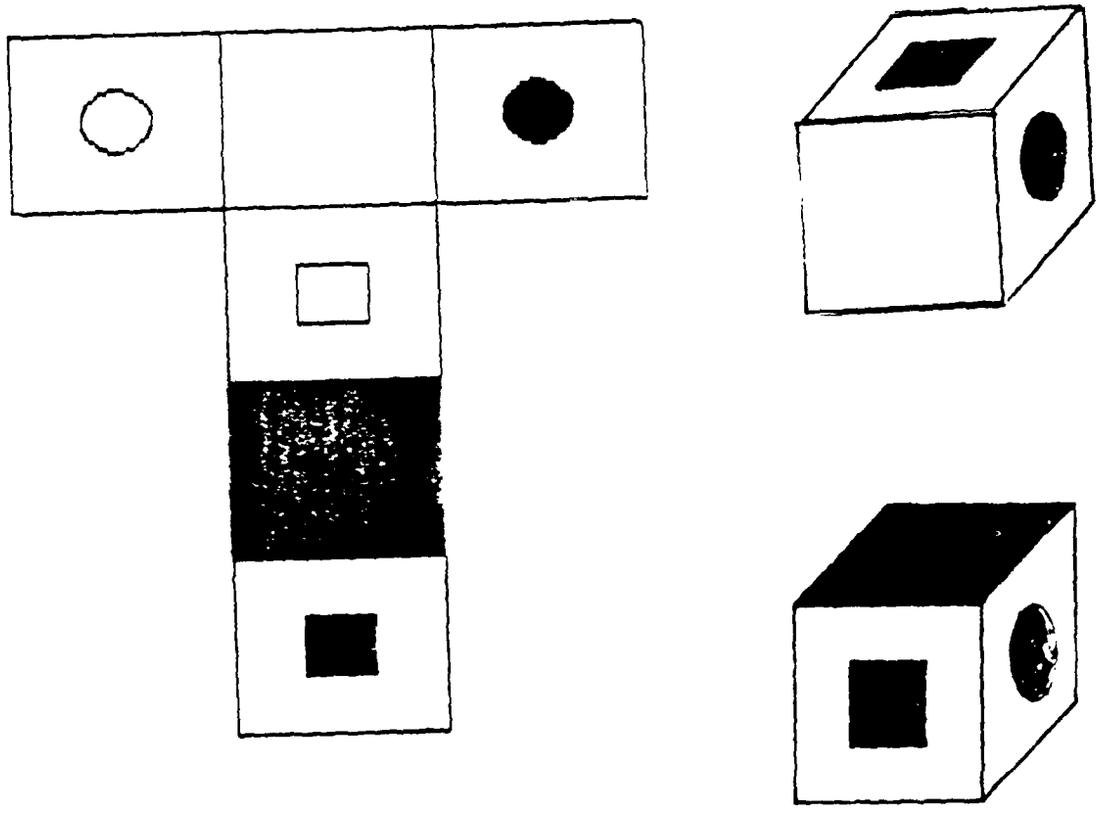


Figure 3-5 presents the mean accuracy for the false problems by condition. The surface effect was not significant, but degrees of rotation had both a significant linear effect ($p<.001$) and a quadratic effect ($p<.01$). The interaction of surfaces and degrees was not significant.

Although the main effects of the attached folding model, the number of surfaces carried and the degrees of rotation, were significant, with effects generally in the expected direction, the significant interactions indicate that the relationships in these data are complex. The main contribution to the interaction is the three surface problems.

An intensive inspection of the three-surface problems revealed that changes in the axis of rotation vary in the three surface problems. That is, after attaching the adjacent pair, the third side may have to be rotated around two separate axes, defined by the traditional coordinates of the cube. Figure 3-6 shows two different correct alternatives that involve carrying three surfaces. For the first alternative, the third side is rotated in only one direction. For the second alternative, the third side is rotated in two directions and thus involves a change of axis or an oblique rotational axis. For the one surface problem, it is not possible to change axis. For the two surface problems, in contrast, the four shapes used here always involve a change of axis.

New Studies Funded in This Project

Modeling Building for the Spatial Folding Task.

The purpose of this study was to test the attached folding model more completely as an explanation of processing on the spatial folding task in several ways. First, an unconfounded test of the independence of the anchoring and folding stage was needed. In the attached folding model, it

is hypothesized that the degrees of rotation and the number of surfaces carried influence separate processing stages (i.e., anchoring and folding, respectively). However, the previous study found a significant interaction of the degrees of rotation with the number of surfaces carried, which indicates dependence between the stages.

An inspection of the tasks that were used in the previous study led to the hypothesis that an uncontrolled variable in the three surface problems, change in the folding axis, may have caused the interaction. Thus, the change of axis variable needs to be controlled to provide a more adequate test of the model. Thus, the present study contained four levels of the surfaces carried variables (1, 2, 3- and 3+), where the latter two levels distinguished between three-surfaces problems without and with a change of axis, respectively. It was predicted that the three-surface problems with the change of axis would be more difficult than those with no change of axis; however, a linear trend for the surfaces variable was no longer expected. Thus, a quadratic trend was hypothesized. As in the previous study, the degrees of rotation effect was hypothesized to be linear.

Second, the empirical plausibility of alternative mental models for the spatial folding task needed to be examined. Two alternative models that have theoretical plausibility are the direct folding model and the verbal analytic model, as was described above. These alternative models could be either empirically superior to the attached folding model, or indistinguishable from it, on the current version of the spatial folding task. A comparison between the models was planned from mathematical modeling of response times and accuracies for the individual items. To provide a sensitive test, stable estimates of the item means were needed.

Thus, a larger sample size was planned.

Third, an assessment of individual differences in mental models for the spatial folding task processing strategies was needed. Other studies on spatial processing (e.g., Just & Carpenter, 1985; Egan, 1979; Cooper, 1983) found that individuals are differentially characterized by the alternative mental models. Although one mental model may provide a better general description of the spatial folding task, individuals may vary in the extent to which they apply each model. Thus, a comparison of the goodness of fit of the three mental models over individuals who report using different strategies was desired.

Method

Materials. The spatial folding stimuli consisted of 147 cube-folding tasks with unique markings on each side. As in the previous study (e.g., Figure 3-1), four stem shapes were used and all tasks were presented in the verification format. All spatial folding tasks were displayed on a CRT. The MICROCAT program (Vail, 1985) was used to construct the items, control the display and record response time and accuracy.

For three stem shapes, three-surface problems can be created with or without a change of axis to fold the third side. On Figure 3-1, these stem shapes are the two on the first row ("T" and "Z") and the stem shape on the lower left ("F"). For the remaining stem shape on the lower right ("X"), all three-surface problems involve a change of axis.

Ninety true problems were constructed to vary in degrees of rotation (0,90,180) and surfaces carried (1,2,3-,3+), for a design of Degrees(3) X Surfaces(4) from the four shapes. Two stimulus sets that differed in the configurations of the marked sides provided additional replications of the

conditions. Eight replications were observed for all conditions except the three-surface problems with no change of axis. For these tasks, only six observations per condition were available, due to the impossibility of constructing three-surfaces problems without a change of axis from the "X" shape.

Fifty-seven false problems were also constructed. An additional level of the surfaces carried variable (zero surfaces) was constructed for problems in which the stem could not be anchored to the folded alternative, due to mismatched markings on the sides. Thus, the design was Degrees(3) X Surfaces(5) from the four shapes. Each shape was observed on only one stimulus set, counterbalanced over conditions. All conditions except those with three-surface problem without a change of axis were observed on four tasks. The conditions with three-surface problems without a change of axis were observed on only three tasks each.

A second set of materials was the self-reported strategy questionnaire. One way to increase the accuracy of self-reported strategies is to minimize the role of memory (Ericsson & Simon, 1983). Thus, to eliminate retrospective reporting of strategies, each statement about strategies was preceded by a spatial folding item to which the strategy could be applied. The strategy statements were compiled from descriptions of spatial problem-solving that were given by students in a small research seminar. Also, a number of general statements about spatial skills were taken from questionnaires in other studies, such as Vandenburg (1985). All items were rated on a 5-point scale, with the options ranging from "Almost Never" to "Nearly Always".

Design. Since the attached folding model postulates that attaching and

folding are accomplished in sequentially dependent stages, the variables that influence these stages, degrees of rotation and number of surfaces carried, respectively, were predicted to have significant and additive effects on response time. Furthermore, consistent with other studies, a linear trend was predicted for the degrees of rotation, while either a linear or quadratic trend was predicted for the number of surfaces carried. Although the attached folding model does not lead to specific expectations for accuracy, consistent with other studies, it was anticipated that both the degrees of rotation and the number of surfaces carried would have significant effects. Within-subjects analyses of variance, with a design of Degrees(3) X Surfaces(4), were planned for both response time and accuracy.

The three mathematical models were operationalized by scoring items on the task stimulus features that were hypothesized to influence the difficulty of specific processing stages. For the attached folding model, three variables were scored. The attachment process was represented by degrees of rotation, which was scored as actual degrees to reflect a hypothesized linear effect. The folding process was scored by two variables, to represent the linear and quadratic components of the surfaces carried variable. That is, the four categories of surfaces carried (ordered as 1,2,3-,3+) were weighted by orthogonal polynomials to represent the linear and quadratic trend components.

For the direct folding model, four variables were scored to represent the folding and rotating processes. The folding process was represented by three variables. Since the folding process depends on the unfolded stem, without respect to any alternative, differences in stem shapes are the only

source of influence on processing difficulty. Thus, three orthongonal contrasts were scored to reflect the four stem shapes. The rotation process was represented by one variable, the amount of rotation required to orient the mentally folded cube (from the stem) to the alternative. Following Just and Carpenter (1985), two versions of the cube comparison process were distinguished; rotating with a standard (rigid) axis, which corresponds to the dimensions of the cube, and rotating with an oblique axis, which defines the most efficient rotation to the target. For the rigid rotation model, the number of 90-degree rotations required to align the cubes was scored. For the oblique rotation model, scoring was identical to the rigid rotation model except that problems with two 90-degree rotations in different directions, were scored as one and a half rotations.

For the verbal-analytic model, two binary variables were scored. One binary variable represented items that could be solved by adjacent relations between the pairs of markings on the sides. The other binary variable represented items that could be solved by position relation of the third side with respect to the adjacent pair. An enhanced verbal-analytic model added two variables to reflect the postulated interaction with the degrees of rotation. That is, additional propositions to further specify (or reverse) the relationships are required when the degrees of rotation are 90 and 180. Thus, the two binary variables for adjacent relations and position relations were multiplied by the degrees of rotation to obtain the interactions.

Comparisons of the models for subjects who may be using different information-strategies were also planned. The strategy self-report

questionnaire, as described above, was to be used to divide subjects into strategy groups. The three models were also compared within strategy groups.

Subjects. The subjects were 83 undergraduates at a large midwestern university who were participating in the experiment as part of a course requirement. Data from an additional 9 subjects were collected, but were lost due to a hard disk failure on one of the microcomputers.

Results

Prior to data analysis, the data were trimmed for low accuracy rates and insufficient response time. Fourteen subjects who did not perform significantly better than expectation from random guessing model were eliminated for low accuracy levels. Additionally, all responses of less than one second were eliminated for inadequate response time to the task. Trimming for low response time eliminated less than one observation per subject.

Means for each subject were calculated over the tasks within each condition. Accuracy was analyzed as a log odds ratio ($\ln(P/1-P)$) to be comparable in scale to an item response model. Response time was calculated as the geometric mean response time over the tasks within a condition. The analyses on response times were performed on total response time, rather than correct response times, due to substantial missing data for the latter. That is, complete correct response time data were available for only 13 subjects, since the error rate on the true problems was substantial (22.5 percent), as expected. Thus, to eliminate sampling bias, total response times are reported. However, when the analyses were repeated on the 13 subjects with complete correct response

time data, and the same pattern of results were obtained.

Analysis of Variance. Table 4-1 presents results from a within-subject analysis of variance of response time for true problems. Table 4-1 presents an analysis for the three stem shapes with all four levels of the Surfaces variable. It can be seen that both degrees of rotation and number of surfaces carried had significant main effects, as expected. However, the additivity condition did not hold, as the Degrees X Surfaces interaction was significant. In Figure 4-1, the mean response times are plotted by conditions. The two most significant departures from additivity are 1) the one-surface problems do not show increased response time until 180 degrees of rotation and 2) the three-surface problems show increased response time from 0 degrees to 90 degrees only. Table 4-2 shows that the same pattern of results were obtained when the "X" shape was inconcluded in the analysis (eliminating the Surfaces 3- level).

Table 4-3 and Table 4-4 show the same pattern of results also applied to accuracy. That is, degrees of rotation and number of surfaces carried had significant main effects, but also significant interactions. Figure 4-2 shows that accuracy decreased sharply from 0 to 90 degrees for the two and three surface problems with a change of axis. The three-surface problems with a change of axis were the most difficult problems, and show little relationship to degrees of rotation. The one-surface problems were the easiest problems, and although accuracy decreases from 0 to 90 degrees, the decrease was not as rapid as for the two and three surface problems with a change of axis.

Table 4-5 shows the analysis of variance on response time for the false problems. As for the true problems, significant main effects for

Table 4-1

Within Subject Analysis of Variance on Response Time

for True Tasks[1]

Source	df	MS	F	P
Degrees of Rotation[2]	3,204	653.45	72.28	.000
Surfaces Carried[3]	2,136	1146.33	106.76	.000
Degrees X Surfaces	6,408	214.21	29.76	.000

[1]Shapes = T, Z, F

[2]Linear Component Only ($F = 164.61$, $p < .001$)

[3]Linear Component ($F = 106.71$, $p < .001$)

Quadratic Component ($F = 5.05$, $p = .028$)

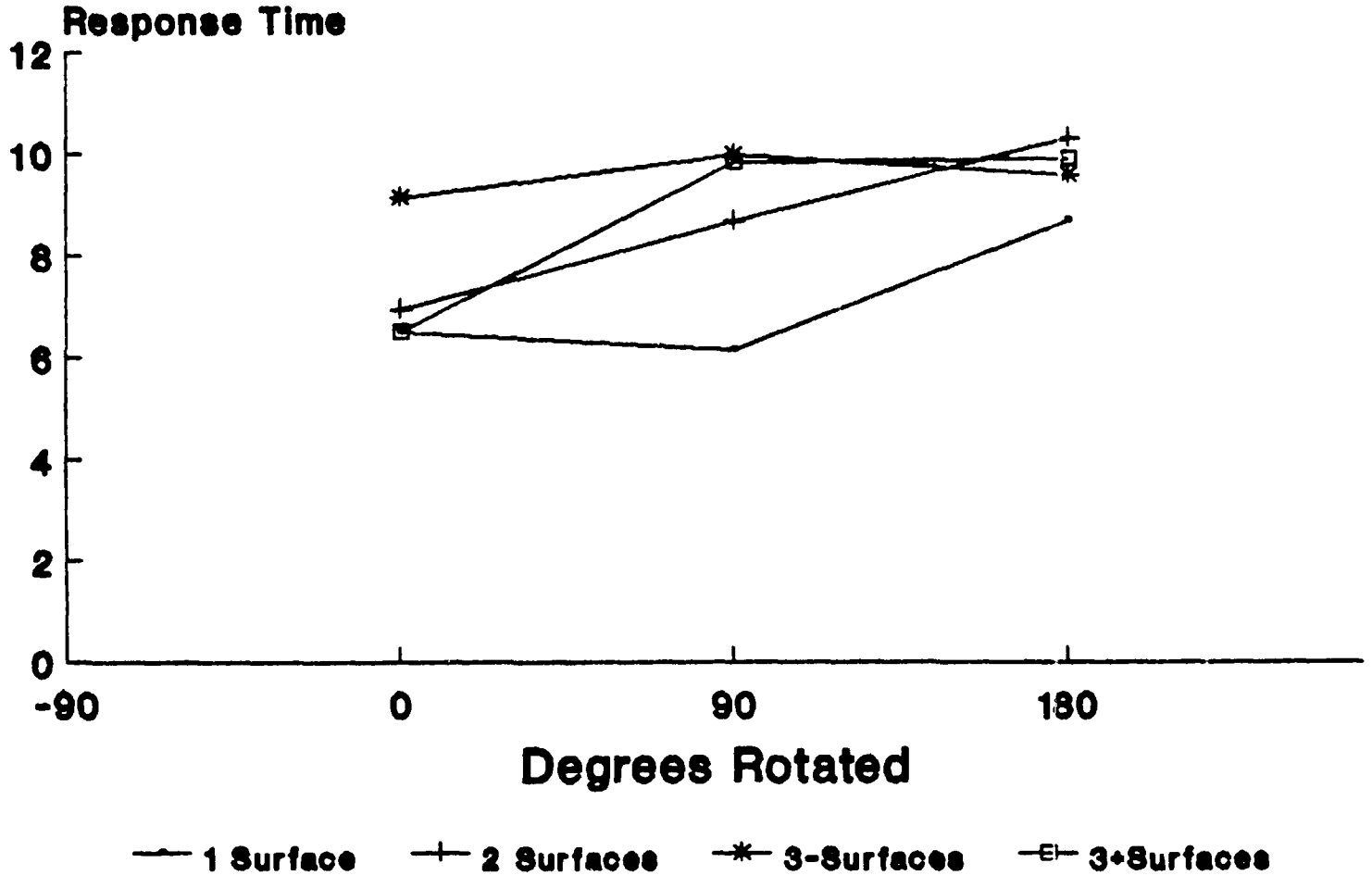

Response Time
12
10
8
6
4
2
0
-90
0
90
180
Degrees Rotated
1 Surface
2 Surfaces
3-Surfaces
3+Surfaces

Table 4-2

Within Subject Analysis of Variance on Response Time for True Tasks[1]

Source	df	MS	F	P
Degrees of Rotation[2]	2,136	1514.47	129.59	.000
Surfaces Carried[3]	2,136	871.58	96.65	.000
Degrees X Surfaces	4,272	191.68	27.72	.000

[1]All Shapes

[2]Linear Component Only ($F = 187.96$, $p < .001$)

[3]Linear Component ($F = 126.97$, $p < .001$)

Quadratic Component ($F = 30.80$, $p = .001$)

Table 4-3

Within Subject Analysis of Variance on Response Time for True Tasks[1]

Source	df	MS	F	P
Degrees of Rotation[2]	2,136	5.78	48.85	.000
Surfaces Carried[3]	3,204	3.00	25.83	.000
Degrees X Surfaces	6,408	.95	10.94	.000

[1]Shapes = T, Z, F

[2]Linear Component Only ($F = 87.79$, $p < .001$)

[3]Linear Component ($F = 30.36$, $p < .001$)

Quadratic Component ($F = 6.05$, $p = .02$)

Table 4-4

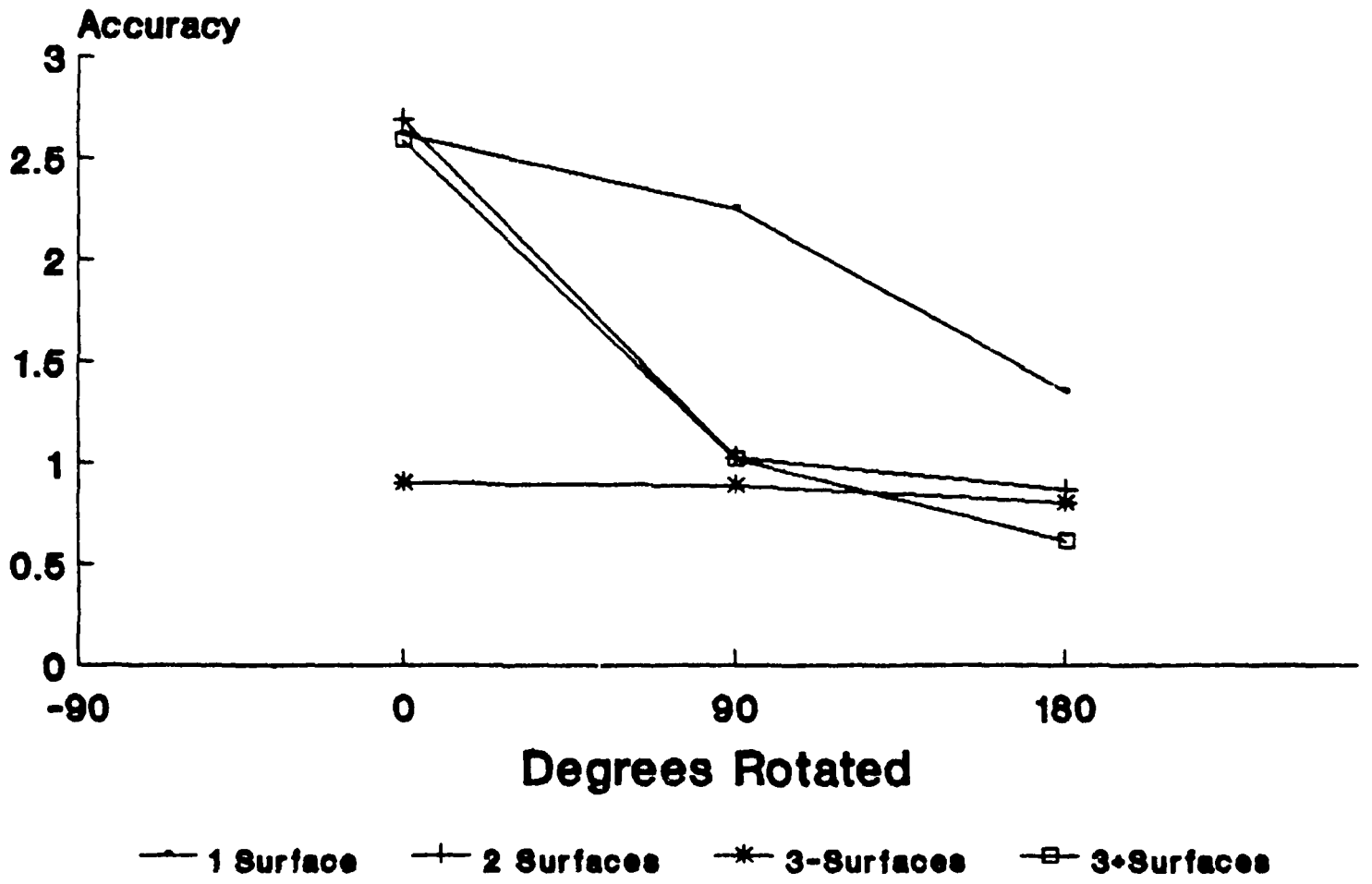
Within Subject Analysis of Variance on Response Time for True Tasks[1]

Source	df	MS	F	P
Degrees of Rotation[2]	2,136	7.58	66.72	.000
Surfaces Carried[3]	2,136	3.81	36.09	.000
Degrees X Surfaces	4,272	.69	7.79	.000

[1]All Shapes

[2]Linear Component Only ($F = 118.81$, $p < .001$)

[3]Linear Component ($F = 66.55$, $p < .001$)



Accuracy is ln(P/1-P)

Table 4-5

Within-Subject Analysis of Variance for Response Time on False Tasks[1]

Contrasts	df	F	P
Anchoring effect	1,492	342.99	.000
Degrees			
Linear Component	1,492	25.50	.000
Quadratic Component	1,492	.01	.926
Surfaces			
Linear Component	1,492	29.70	.000
Quadratic Component	1,492	.61	.438
Cubic Component	1,492	2.16	.148
Degrees X Surfaces			
Linear (D) X Linear (S)	1,492	8.35	.006
Linear (D) X Quad (S)	1,492	.84	.363
Linear (D) X Cubic (S)	1,492	12.32	.001
Quadratic (D) X Linear (S)	1,492	1.29	.260
Quadratic (D) X Cubic (S)	1,492	36.58	.000

[1]Shapes = T, Z, F

N = 42

both Surfaces and Degrees were found, and the interaction was also significant. Also, similar to the true problems, Degrees had a linear trend. However, unlike the true problems, only the linear component was significant for the Surfaces variable. Also, problems that could be falsified because the stem could not be anchored to the alternative required significantly less processing time than the other conditions.

Model Comparisons. The three information-processing strategy models were examined by regressing response time and accuracy on the independent variables, as defined above. The dependent variables were means on the true items, averaged over the replications on the two stimulus sets. Table 4-6 presents results on the goodness of fit of the three models for response times. For each model, a hierarchical regression analysis was performed. The variables were stepped into the equation in the order that is given on Table 4-6.

It can be seen that the attached folding model, without any terms to represent the interaction of Surfaces and Degrees, achieves relatively good fit, as indicated by the multiple correlation coefficient of .75. It can be seen that adding stem shape to the model did not significantly increase fit. Furthermore, both rotating and folding had substantial contributions to the model, as the proportion of task variance explained by them is nearly equal (.29 and .28, respectively).

Table 4-6 also shows that the simple verbal-analytic model, without variables to represent propositional changes for 90 or 180 degrees of rotation, significantly predicted response time (R = .51). When the propositional change variables were added to the model, the goodness of fit was significantly and substantially increased. In fact, the multiple

Table 4-6

Mathematical Models of Response Time in Spatial Problem Solving Tasks

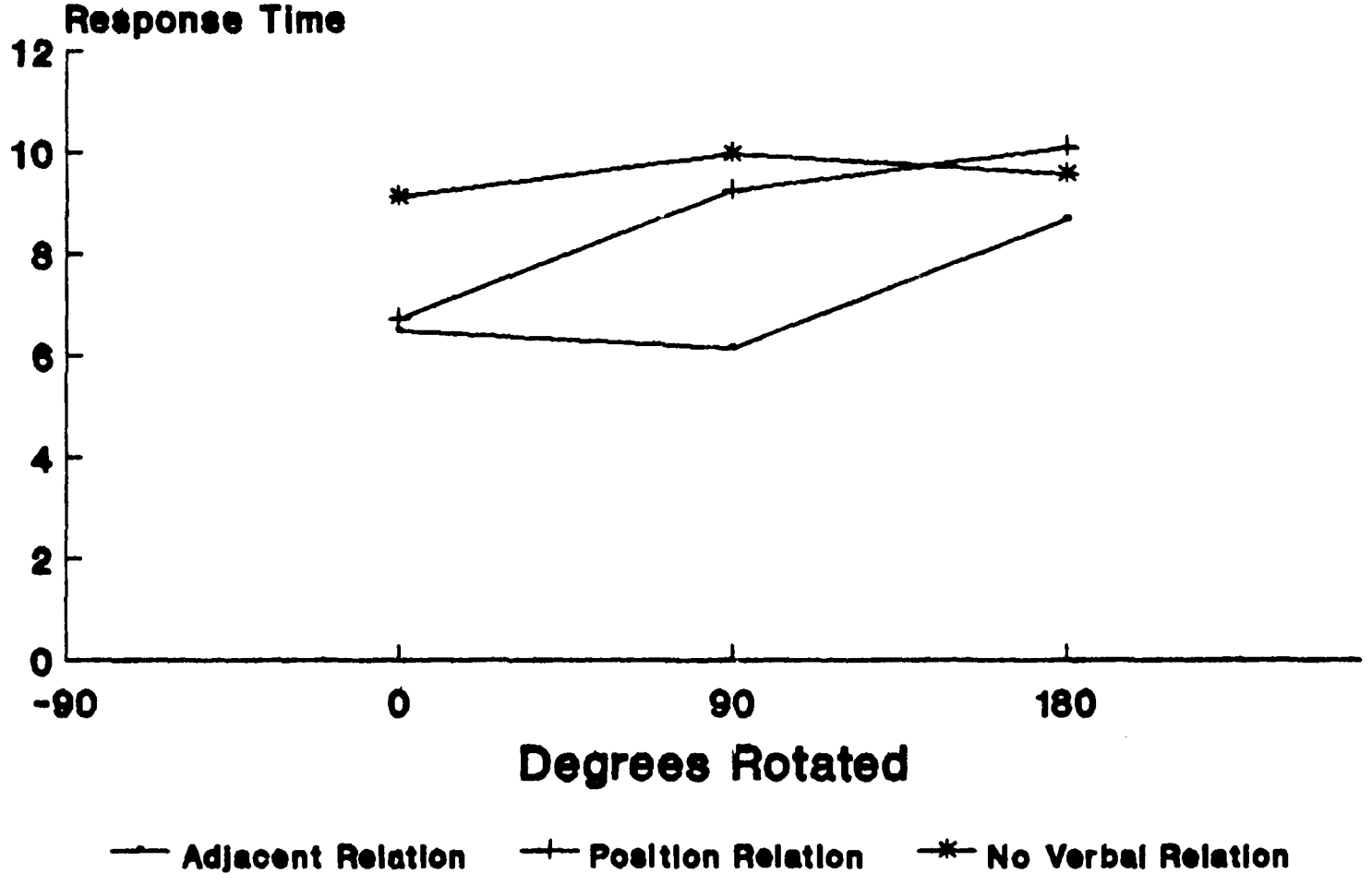
			Model		Change	
Model	df	R	R2	F	R2	F
Verbal-Analytic						
Simple	2	.51	.26	7.53**	-	-
Enhanced	4	.79	.62	16.36**	.36	18.95
Enhanced + Shape	7	.82	.67	10.65**	.05	1.86
Attached Folding						
Rotating	1	.54	.29	17.99**	-	-
Rotating + Folding	3	.75	.57	17.83**	.28	13.34**
Rotating, Folding + Shape	6	.78	.61	9.89**	.04	1.52
Direct Folding						
Folding	3	.15	.02	.30	-	-
Folding, Rotating (Rigid Axis)	4	.54	.29	4.11**	.27	15.21**
Folding, Rotating (Oblique Axis)	4	.52	.27	3.71**	.25	13.69**

correlation coefficient slightly exceeded the correlation coefficient for the attached folding model. However, as for the attached folding model, adding shape to the model did not significantly increase fit.

Figure 4-3 plots the means for tasks with adjacent relations, position relations and no verbal relations by degrees of rotation. It can be seen that the adjacent relation problems increased in processing duration only for the 180 rotations. For problems with position relations, the increase with degrees of rotation appeared linear, while for problems with no verbal relations, the increased processing duration occurred from 0 degrees to 90 degrees.

Table 4-6 shows that the direct folding model, both with rigid and oblique axis, provided substantially worse fit to item response times than the attached folding model or the verbal-analytic model. It can be seen that the highest multiple correlation obtained is .54 for the rigid-axis model. The model with an oblique axis of rotation had slightly worse fit. Also, the sequential analysis indicated that stem shape did not significantly predict response time. Thus, the folding of the shapes was not associated with differential processing difficulty. Figure 4-4 plots the mean response times for the various cube rotations for the direct folding model. It can be seen that response time increased primarily when two rotations, in any direction, were required.

Table 4-7 presents the results from the mathematical modeling of task accuracy. The pattern of results were similar to the results on response time. That is, although the attached folding model, with no interaction terms, provided significant and substantial prediction of task accuracy, the attached folding model was not clearly superior to the verbal-analytic


Response Time
12
10
8
6
4
2
0
-90
0
90
180
Degrees Rotated
Adjacent Relation
Position Relation
No Verbal Relation

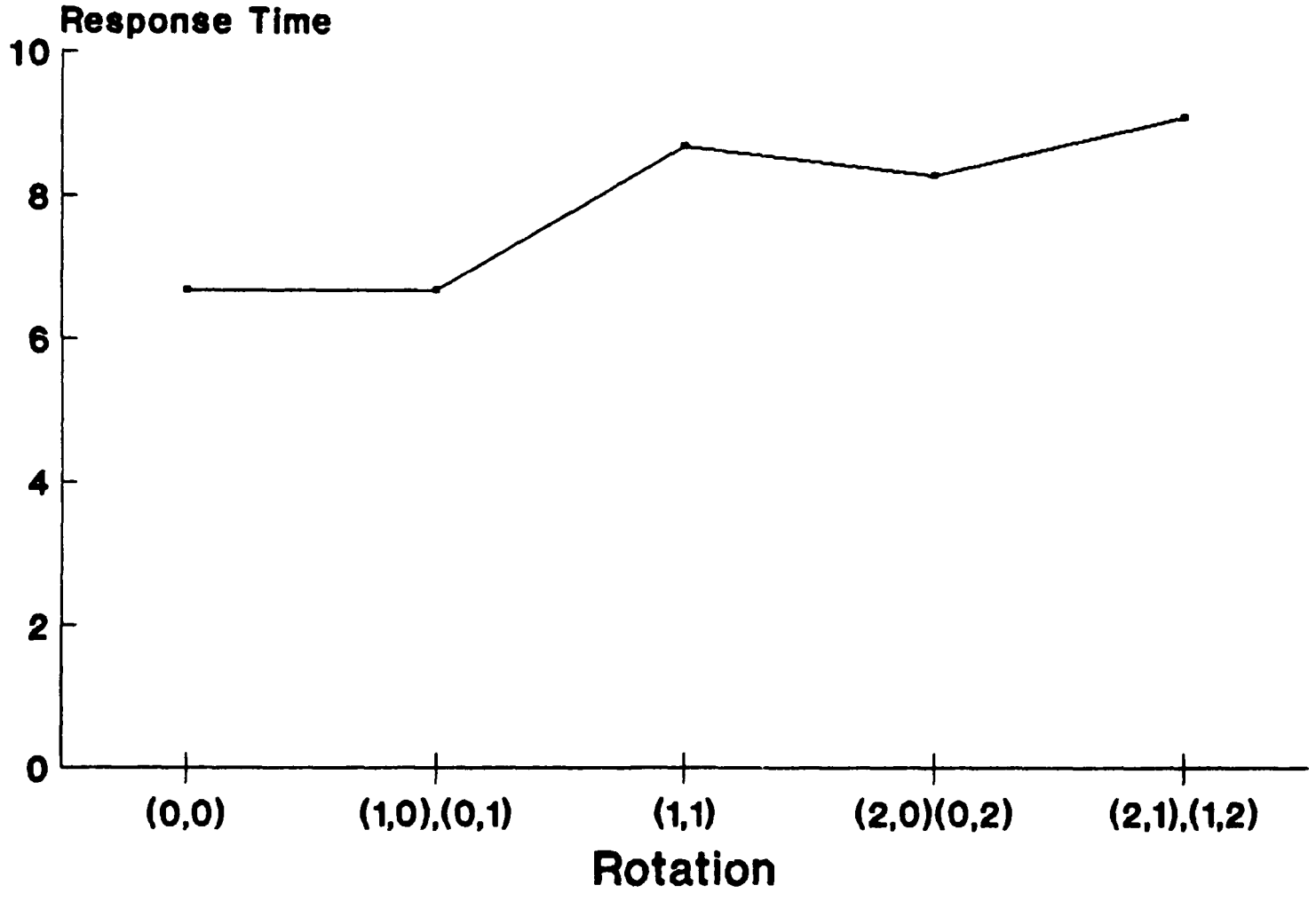
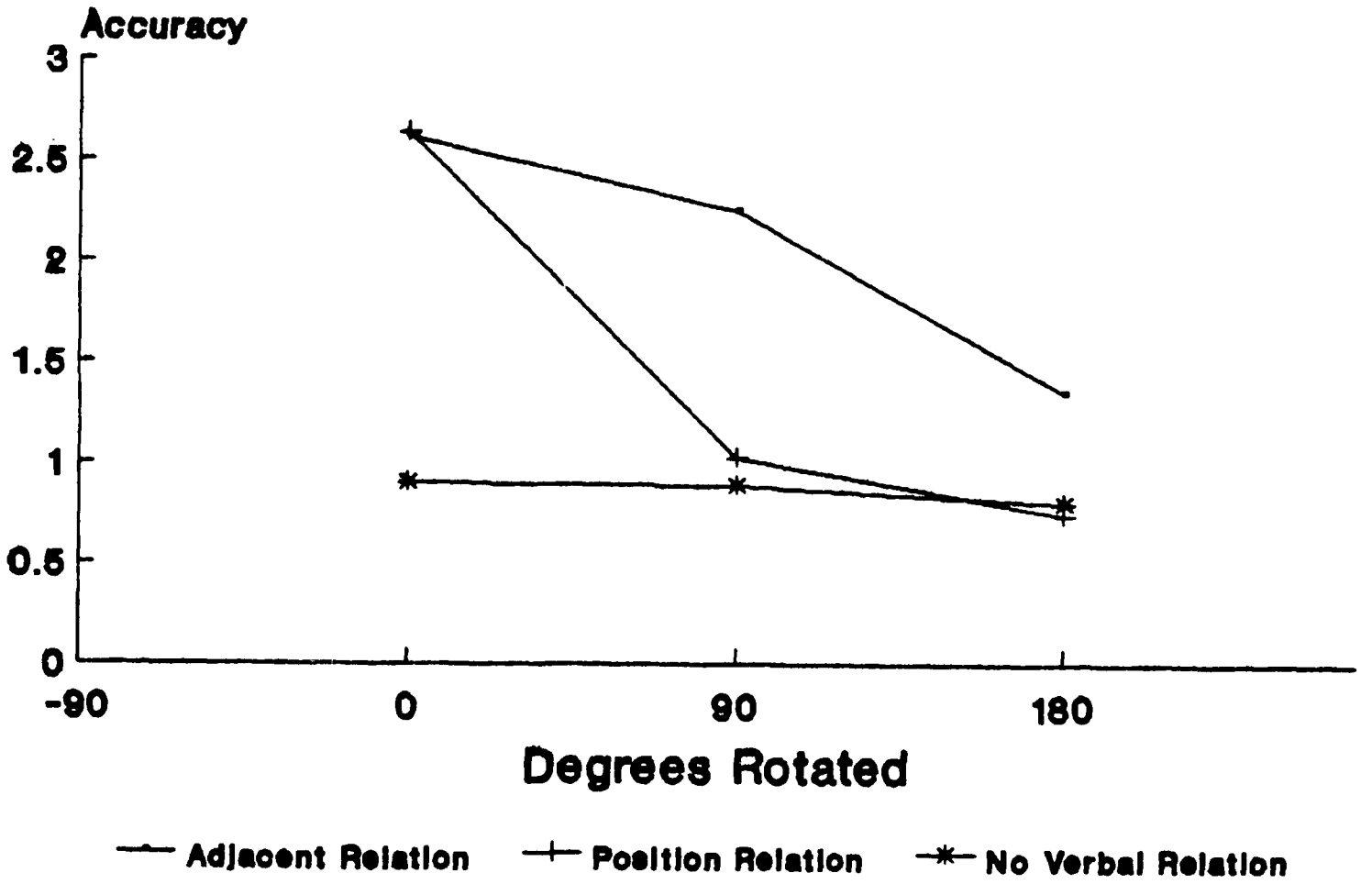

Response Time
10
8
6
4
2
0
(0,0)
(1,0),(0,1)
(1,1)
(2,0)(0,2)
(2,1),(1,2)
Rotation

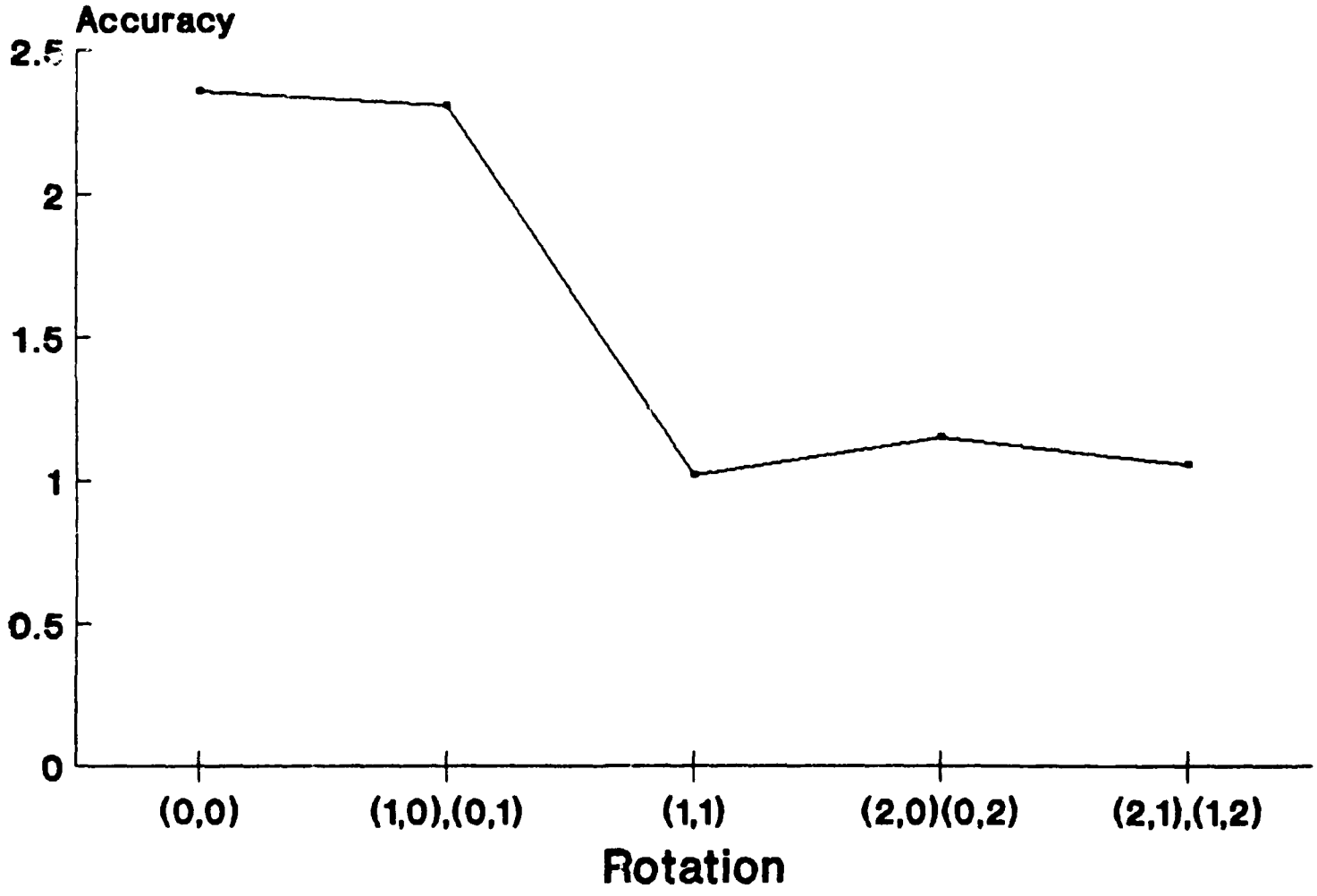
Table 4-7

Mathematical Models of Task Accuracy In Spatial Problem Solving Tasks

			Model		Change	
Model	df	R	R2	F	R2	F
Verbal-Analytic						
Simple	2	.51	.26	7.53**	-	-
Enhanced Verbal	4	.79	.62	16.36**	.36	18.94**
Enhanced + Shape	7	.87	.76	16.57**	.08	4.11
Attached Folding						
Rotating	1	.60	.35	23.69**	-	-
Rotating + Folding	3	.77	.60	20.55**	.25	7.81**
Rotating, Folding + Shape	6	.81	.66	12.32**	.06	1.12
Direct Folding						
Folding	3	.19	.04	.51	-	-
Folding Rotating (Rigid Axis)	4	.60	.36	5.66**	.34	21.25**
Folding Rotating (Oblique Axis)	4	.54	.29	4.17**	.25	14.08**



Accuracy is ln(P/1-P)



Accuracy is ln(P/1-P)

model. In fact, the verbal-analytic model had a slightly higher multiple correlation. It should be noted that the model contained an additional predictor. As for the response time data, the direct folding model did not fit the data as well as the other two models.

Figure 4-5 plots the mean accuracies for the independent variables of the verbal-analytic model. It can be seen that accuracy on both types of verbal problems decreases with increasing degrees of rotation, but position relations problems decreased the most. Problems with no verbal relations changed little with degrees of rotation. Figure 4-6 plots the mean accuracies for the various cube rotations in the direct folding model. It can be seen that the largest decreases occurred from one to two rotations, in any direction.

Self-Reported Strategies. A confirmatory factor analysis of the self-reported strategies for the spatial folding task yielded a model that fit the data (X^2 = 45.11, p=.383). The two a priori factors were Spatial Strategy and Verbal Strategy. Table 4-8 presents the factor loadings.

Factor scores were prepared from the two factors. Figure 4-7 plots factor scores for the subjects. It can be seen that the scores were correlated positively. A 45 degree line from the origin of the plot would define equal strategy scores on the two factors. An error band of .65 was plotted around the 45 degree line. Subjects were classified into one of three strategy groups, 1) Verbal Strategy, the subjects in the lower right area, 2) Spatial Strategy, the subjects in the upper left area, and 3) Combination Strategy, the subjects with nearly equal strategy scores in the middle area. Two subjects with very low scores on both strategy factors were eliminated from further analysis.

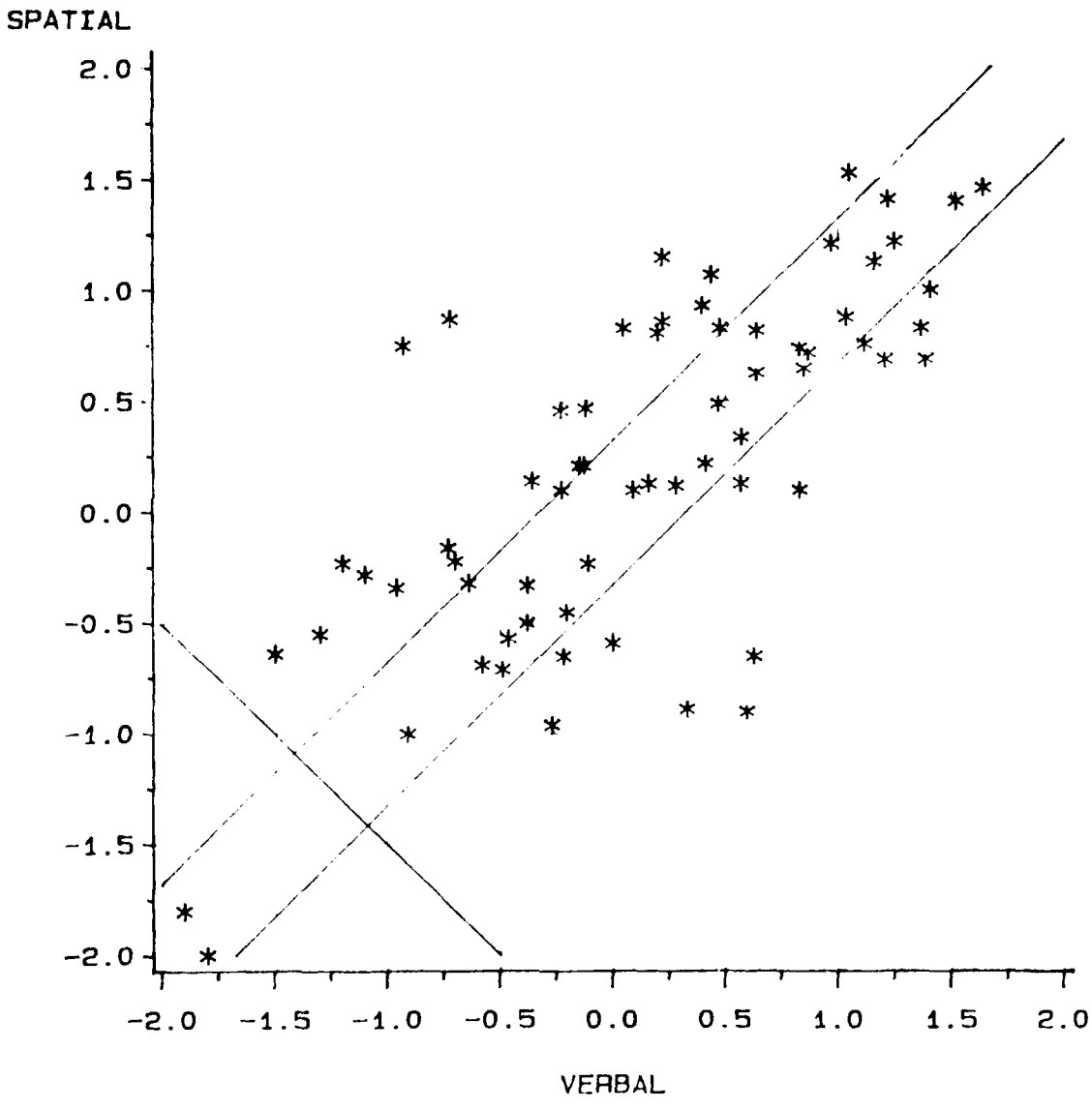
Table 4-8

Factor Loading from Confirmatory Factor Analysis

of Self-Reported Strategies

		Spatial	Verbal
		FI	FII
1)	"I reason about the relative position of the markers."	.000	.609
2)	"I mentally fold the unfolded cube before evaluating the folded cube."	.199	.000
3)	"I reason about the spacing of the markers."	.000	.596
4)	"I match the marker on the unfolded cube, and fold [the stem] around it."	.462	.000
5)	"I rotate the unfolded cube on front of the folded cube, as on an axis."	.622	.000
6)	"I rotate to position markings, then fold."	.531	.000
7)	"I use positional relationships."	.000	.719
8)	"I notice adjacent pairs."	.363	.000
9)	"I fold before rotating."	.537	.000
10)	"I fold and compare."	.645	.000
11)	"I mentally align the markers from the folded cube by rotating it."		

$X^2_{43} = 45.11$, $p = .384$



ALL REASONABLE SUBJECTS N=69

Table 4-9

Descriptive Statistics for Response

and Accuracy by Self-Reported Strategies

		Response Time		Task Accuracy	
Strategy Group	N	Mean	Std	Mean	Std
Verbal	13	8.94	1.93	.75	.13
Combination	27	7.79	1.62	.80	.09
Spatial	23	9.33	1.85	.76	.15

Table 4-9 presents the mean response times and accuracies for subjects in the three self-reported strategy groups. It can be seen that the highest accuracies and the lowest response times were observed for subjects in the Combination Strategy group.

The mathematical models of response time and accuracy were repeated separately for each group. The same pattern of relative goodness of fit was observed within each group as for the total data. That is, dividing into self-reported strategy groups had little influence on the relative goodness of fit between the processing models, or on the specific parameters within any mathematical model.

Discussion

The results clearly indicated that the attached folding model did not provide a better account of the response time and accuracy data when the change of rotational axis variable was controlled. The degrees of rotation and the number of surfaces carried had significant effects, but as in the previous study, the significant interaction did not support independent stages, as postulated in the model. Further, comparisons to alternative mental models indicated that although the attached folding model provided a better account of the data than the direct folding model, it was not better than the verbal-analytic model. Last, dividing subjects into groups according to self-reported strategies did not yield differential fit of the models.

Taken together, the results suggest that the attached folding model is poorly distinguished empirically from a verbal-analytic model in these data. A careful consideration of how the independent variables of the two models were scored for the spatial folding task in this study suggests an

explanation. That is, all problems except the three-surface problems without a change of rotational axis, can be effectively solved without completely folding the third side. For the one-surface problem, propositions about the relative location of the markers can easily be developed so that a verbal analytic strategy was readily available for these problems. For two- and three-surface problems with a change of axis, position relations could be used to solve the item without folding the third side. That is, because the markings on the third side appear the same from all directions, the individual did not have to completely carry the third side into position to provide a confirmation. In Figure 1-1, for example, the relationship that the "white circle marking is to the right of the adjacent pair (the sides marked by the black square and the black side, respectively)" is sufficient to confirm the correct answer (#3). Thus, the spatial folding tasks with non directed markings on the sides did not require completing the folding process.

Thus, these results warrant two conclusions. First, the role of spatial processing is questionable on the variant of the spatial folding task that was used in this study (i.e., a task in which the markings on the sides are non-directed). This has an important implication for the measurement of spatial ability, more generally, because spatial folding tasks with non-directed markings appear quite frequently on spatial ability tests (e.g., Spatial Relations Test on the DAT). That is, high ability scores may indicate either effective spatial processing or effective verbal processing. Second, measuring spatial processing requires a task in which the spatial processing models (i.e., the attached folding model and the direct folding model) are clearly distinguished from a verbal-analytic

processing model. That is, if the majority of items can be solved readily by either strategy, then the models are poorly distinguished. This suggests that a different variant of the spatial folding task may be required to measure spatial processing.

Spatial Folding with Directed Markers

In this study, a new set of items, with directed markings on the sides, was developed. With directed markings, it is possible to construct items that require that the third side is fully folded. Figure 5-1 shows a sample item. In this item, confirming the key requires that the orientation of the markings on the third side is fully known. Three different types of problems can be constructed with directed markings on the sides: 1) adjacent relation problems, in which a propositions about adjacent relations between the markings is sufficient (i.e., all items with only one surface carried), 2) spatial problems, in which the items required spatial processing to confirm the third side (i.e., two and three surface problems) and 3) positional problems, in which a proposition about the relationship of the third side to the adjacent pair is sufficient to confirm the correct answer (i.e., a special subset of two and three surface problems).

Another goal of this study was to examine the impact of the distractor set on processing. Although test developers often speculate that the distractors define the processing required to solve a multiple choice problem, rigorous experimental data often has not been obtained. That is, for problems with the same relationship of the stem to the correct answer, what is the impact of the distractors on processing difficulty? Two competing general hypotheses may be offered. First, the distractor set

may increase processing duration and decrease accuracy, in proportion to the processing required to falsify each distractor. Second, the distractor set may interact with the processes that are applied to confirm the correct answer.

A related goal was to examine the empirical plausibility of the attached folding task when presented in the same format as on ability test items. The attached folding model has been supported experimentally only on verification tasks in the preceding studies. However, psychometric items employ the multiple choice format, rather than the verification format, to minimize the impact of guessing. Thus, the empirical plausibility of the attached folding model for a spatial ability test depends on its success in the multiple choice format.

To maintain the distinction between adjacent relations, spatial and position relations problems, the distractors must be consistent in type with the correct answer. That is, the type of processing that is involved in confirming the correct answer also must be involved in falsifying the distractors. This constraint restricts the variations of the distractor sets in a particular problem type. Thus, distractor sets were varied within problem types.

For the spatial items, the attached folding model provided an explicit rationale for each distractor. In one distractor set, the orientation of the adjacent pair on all distractors was matched to the correct answer. The attached folding model would predict this set to require only repeated confirmation processing, as once the stem is folded to any alternative, the direction of the marking on the mentally folded third side may be compared to each alternative with no further rotation or

folding. In a second distractor set, the orientation of the adjacent pair was rotated in the distractors. The attached folding model would predict more processing for this set, because additional rotation is required. In a third distractor set, the distractors vary in both the orientation of the distractors and in the adjacent pair that is anchored. The attached folding model would predict greater processing for this distractor set, since the stem is anchored to each alternative.

For adjacent relation problems, the verbal analytic model generates predictions about distractors. Distractor sets that have the same adjacent pair with the same orientation do not require any reversal or changes in propositions about the markings that could be generated from the stem relationships. Thus, this distractor set is predicted to be relatively easy compared to a distractor set in the alternatives change orientations. In the latter case, the propositions about the markings must be reversed or changed to falsify the distractors.

For the positional problems, either the verbal analytic or attached folding model may generate predictions about distractor difficulty, as either strategy could be applied to solve these problems. However, the definition of the positional problems, precludes constructing more than one distractor with the adjacent pair in the same orientation as the correct answer. Thus, a set of three distractors with the same orientation could not be constructed for the positional problems. However, it was possible to construct a set in which the distractors varied in orientation, but had the same adjacent pair. Further, it was also possible to construct distractors that varied in both orientation and the adjacent pair. Last, a special distractor set with paired distractors was also

constructed. This will be described more thoroughly in the methods section.

Method

Subjects. The subject were 55 undergraduates who participated in the experiment to fulfill a course requirement. The data from four subjects were lost due to an equipment failure.

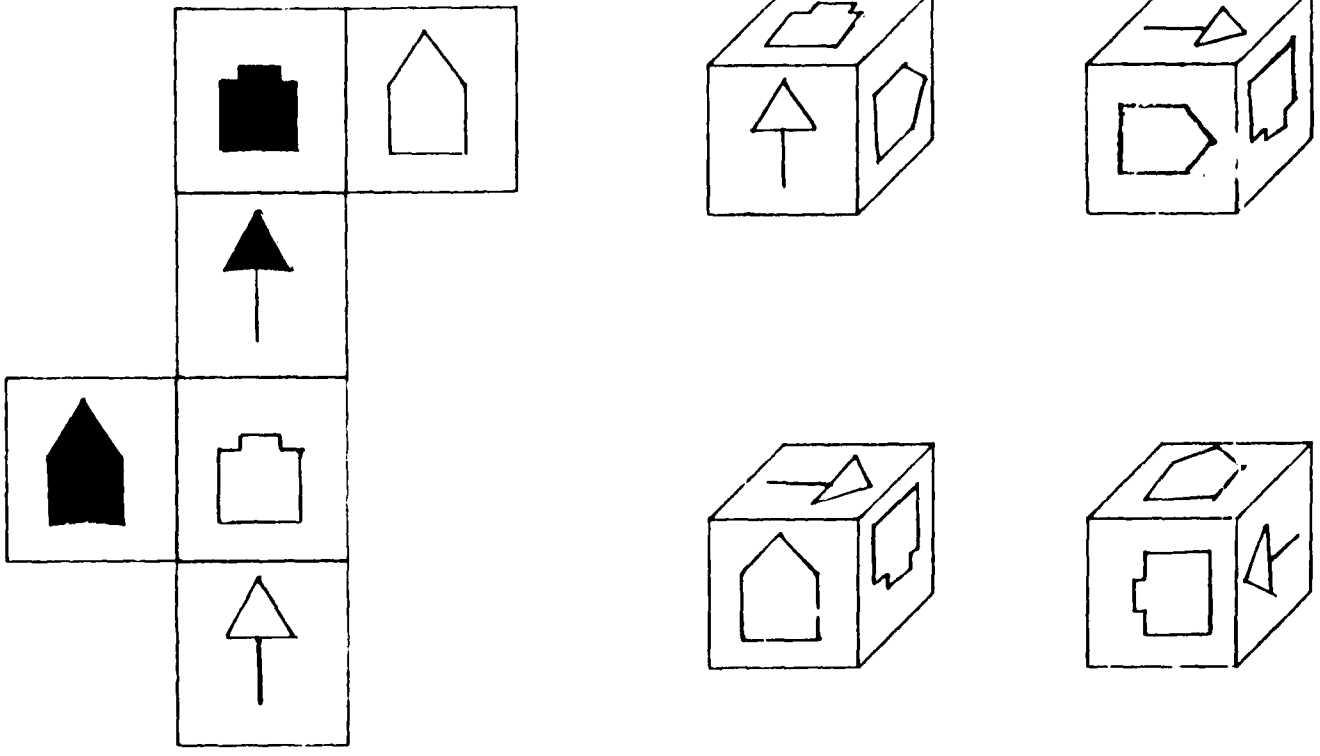
Procedure. All subjects were presented the full item set on a microcomputer. Commercially available software (MICROCAT) was used to present the items and record response time and accuracy.

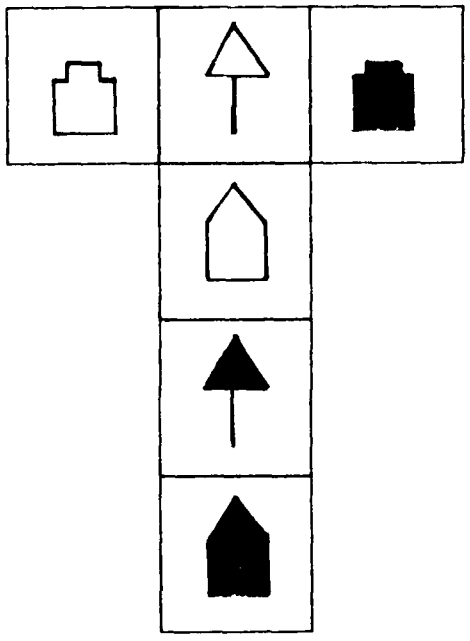
Materials. The spatial folding problems consisted of an unfolded stem with directed markings on the sides and four alternatives as shown on Figure 5-1. Three different stem shapes, shown on Figure 5-2, were used. The configuration of the markings was counterbalanced with stem shape, so that every item was unique. A total of 99 test problems were constructed.

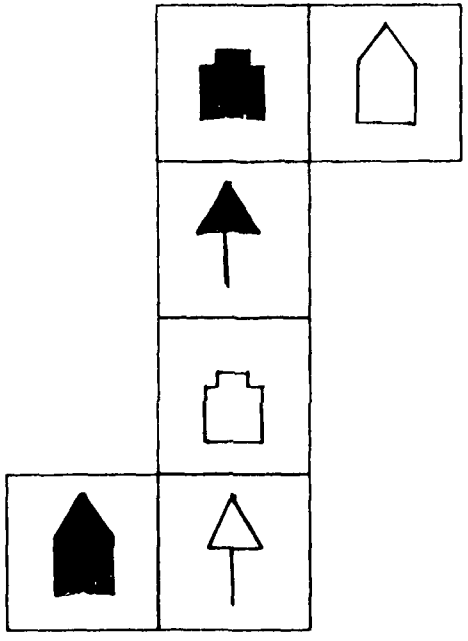
Design. For all three problem types, the relationship of the correct answer to the stem was varied by the degrees of rotation. Figure 5-3 shows the degree of rotation for the correct answer in two surface and three surface problems.

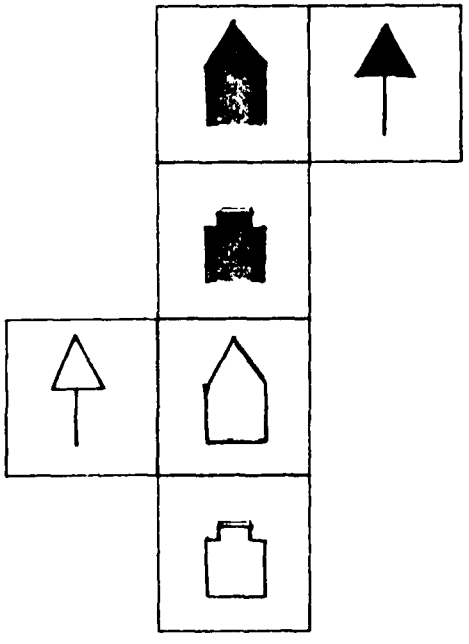
For spatial problems, the degrees of rotation and the number of surfaces carried to confirm the correct answer was crossed with type of distractor set. Thus, the spatial problems were designed as Degrees(3) X Surfaces(2) X Distractors(3), with three replications each, for a total of 54 problems. Figure 5-4 shows the three types of distractor sets for a problem in which confirming the correct answer requires zero degrees of rotation and two surfaces carried.

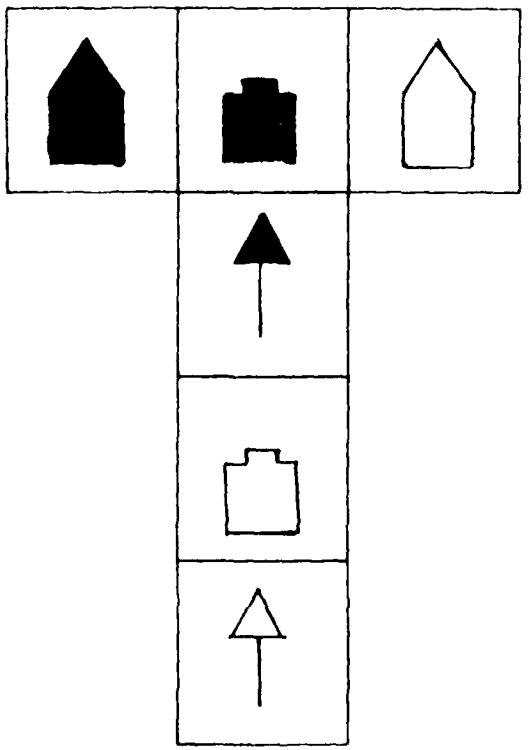
For the other two problem types, fewer factors were varied, due to





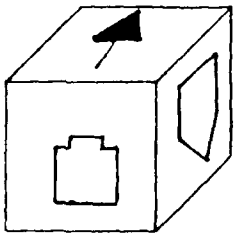




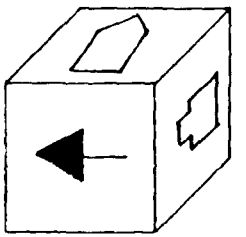


CORRECT FOLDED VIEWS

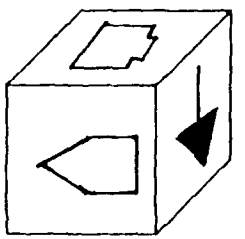
2 SURFACES



0

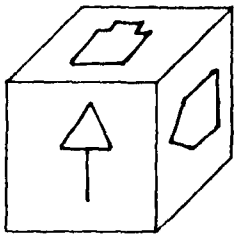


90

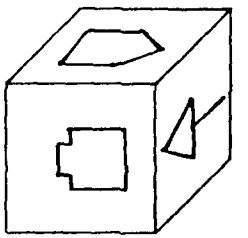


180

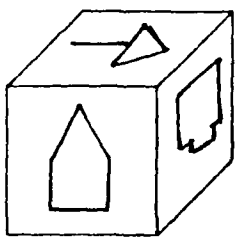
3 SURFACES



0



90

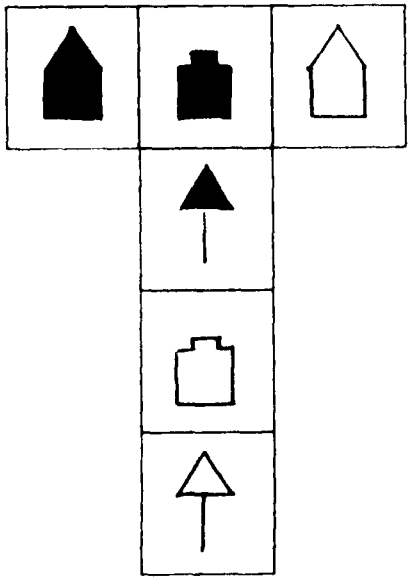


180

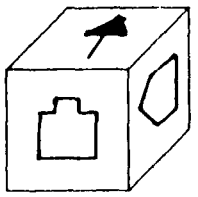
both problem constraints and practical constraints of the testing situation. That is, it was desirable to minimize the number of varied factors to shorten the testing time required for a subject.

For the adjacent relation problems, only degrees of rotation was crossed with distractor set, as all these problems require only one surface to be carried. Furthermore, only two distractor sets , same orientation or rotated orientation as shown on Figure 5-5, were developed. Thus, the adjacent relation problems were designed as Degrees(3) X Distractors(2), with three replications each, for a total of 18 problems.

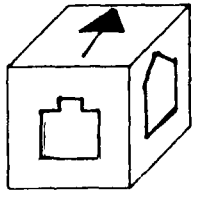
For the position relation problems, only the items in which confirming the correct answer required two surfaces to be carried were developed. Three distractor sets were developed. Rotated orientations, shown on Figure 5-6, were similar to the other problem types, but the other two distractor sets were unlike the other types. The Different Position distractors were constructed according to the same definition as used for the spatial items, as they varied in both orientation and anchoring points. However, the only possible three-surface problems that could be falsified by a position relationship, also could be falsified by a proposition about adjacent relations. Thus, the Different Position distractor sets were hypothesized to be relatively easy. In the Paired Distractors, one distractor had the same orientation of the adjacent pair as the correct answer, so that falsifying the position relationship of the distractor could lead to confirming the correct answer. In summary, the position relation problems were designed as Degrees(3) X Distractors(3), with three replications each, for a total of 27 problems.



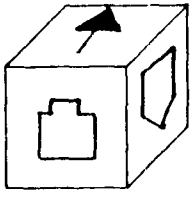
SAME ORIENTATION



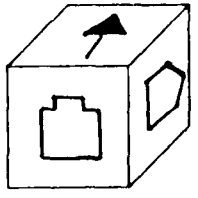
0



0

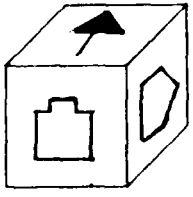


KEY 0

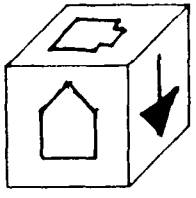


0

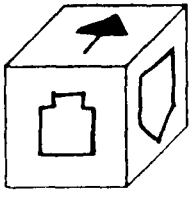
ROTATED ORIENTATIONS



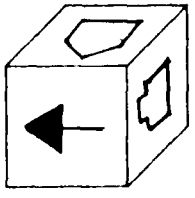
0



180

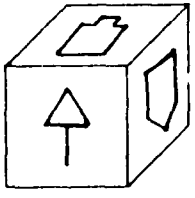


KEY 0

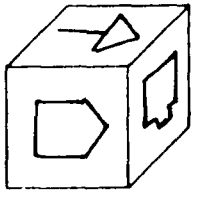


90

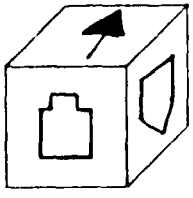
DIFFERENT POSITIONS



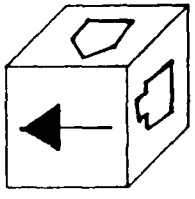
3 SURFACES



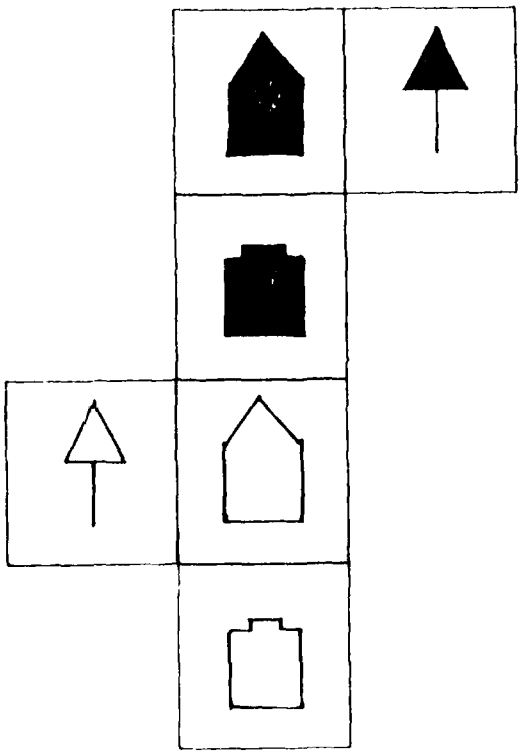
3 SURFACES



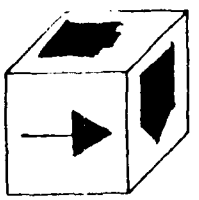
2 SURFACES
KEY



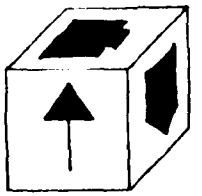
2 SURFACES



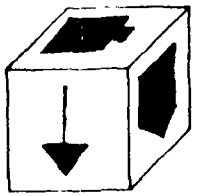
SAME ORIENTATION



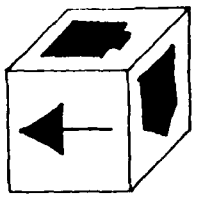
180



180

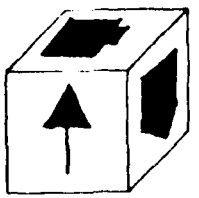


KEY 180

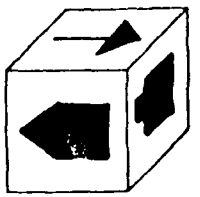


180

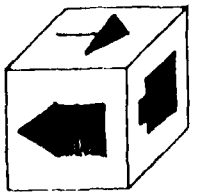
ROTATED ORIENTATION



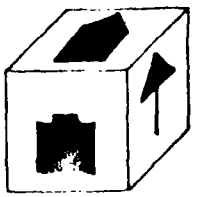
180



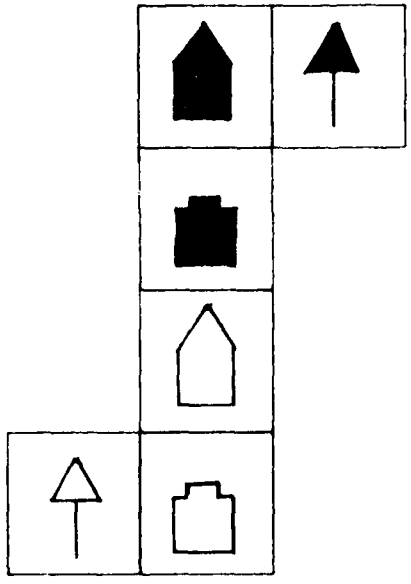
90



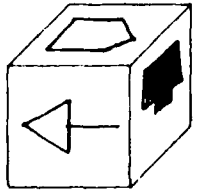
KEY 90



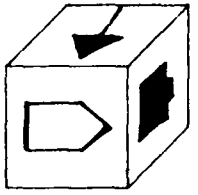
0



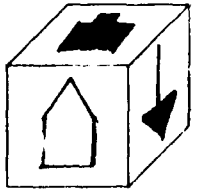
ROTATED ORIENTATIONS



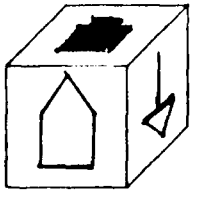
180



90

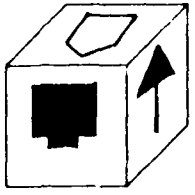


KEY 0

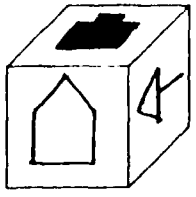


0

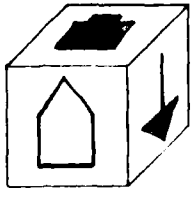
PAIRED DISTRACTORS



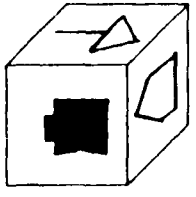
180



0

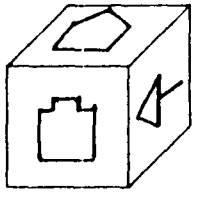


KEY 0

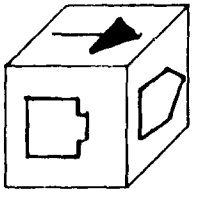


90

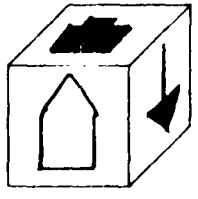
DIFFERENT POSITIONS



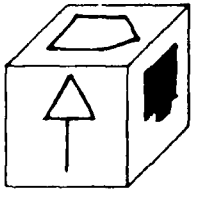
3 SURFACES
0



3 SURFACES
90



2 SURFACES
KEY



2 SURFACES
180

Results

Prior to analysis, the data were trimmed for low accuracy levels. Subjects were eliminated if their accuracy scores were not significantly better than the accuracy rate of .25 that is expected under random choice from among the four response alternatives. Using a one-tailed significance level of .05, eleven subjects were eliminated because their accuracy rates did not exceed the critical value for accuracy rate of .335. For the remaining 40 subjects, at least one of three problems for each condition was correct, so that correct response times could be analyzed.

Response time. The dependent variable was the geometric mean response time for correct items. Incorrect items were treated as missing data for the geometric means. All response times of less than one second were treated as missing data.

The overall impact of the design variables, problem type and distractor set (nested within problem type), was examined by a within subjects analysis of variance. To perform this analysis, the data were collapsed over degrees of rotation, number of surfaces carried (applicable to the spatial problems only) and stem shape.

Table 5-1 presents the mean squares, F values and probabilities for seven orthogonal contrasts that reflect the nested design. For problem type, the contrast of the spatial and position relation problems with the adjacent relation problems was highly significant ($p<.000$). Further, the spatial problems differed significantly from the positional problems ($p<.000$).

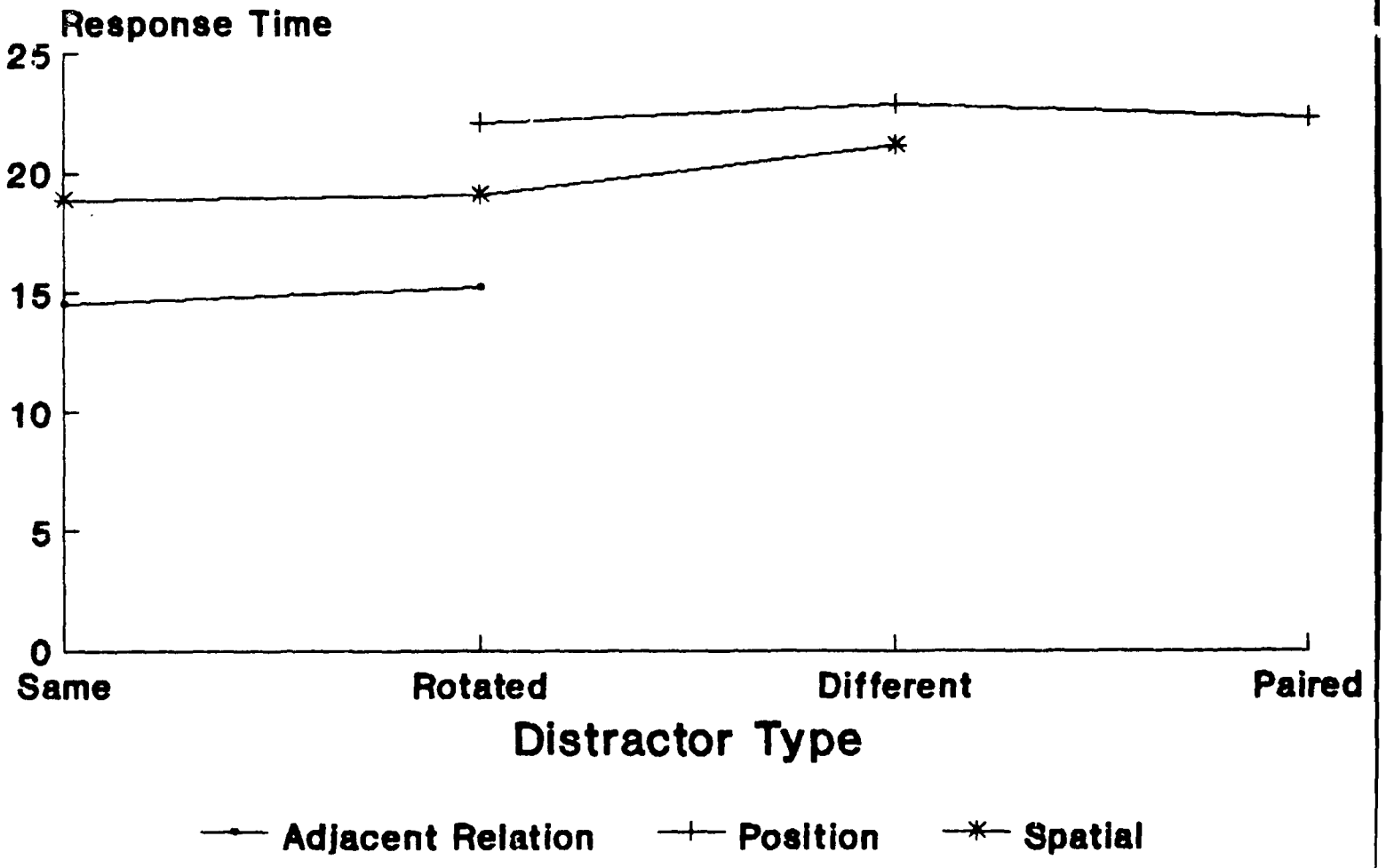
For the distractor set comparisons within problem types, the only significant contrast was within the spatial problems. Here, problems with

Table 5-1

Within Subject Analysis of Variance for

Response Time By Problem Type and Distractor Set

Source	MS	F	P
Problem Type			
Spatial and Position vs. Adjacent Relation	2292.62	141.42	.000
Spatial vs. Position	45.64	17.34	.000
Distractors Within Problems			
Spatial			
Different vs. Same and Rotated	132.20	11.23	.002
Same vs. Rotated	1.20	.10	.752
Adjacent Relation			
Same vs. Rotated	10.48	1.05	.310
Position			
Different vs. Rotated and Paired	11.00	.48	.493
Rotated vs. Paired	.93	.07	.791


Response Time
25
20
15
10
5
0
Same
Rotated
Different
Paired
Distractor Type
Adjacent Relation
Position
Spatial

distractors that had the same anchoring point, presented either in the same orientation or in a rotated orientation, differed significantly from problems with different anchoring points. However, no significant differences in processing durations was observed due to the rotation of the same anchoring pair for spatial items. Further, the distractor sets did not differ significantly for the other problem types.

Figure 5-7 presents the means over problems and distractor sets. It can be seen that the adjacent relation problems had the shortest processing durations, while the position relation problems had the longest processing duration. Also, processing time was relatively longer for spatial problems in which the anchoring point varied between alternatives.

All subsequent analyses of variance on response times were conducted separately within problem types, due to the design constraints that differed between the problem types. Table 5-2 presents a within subjects analysis of variance of Degrees(3) X Surfaces(2) X Distractors(3) for the spatial items. It can be seen that the degrees of rotation had a highly significant main effect. Further, this effect is well described as linear, because the linear trend component was highly significant ($p=.000$) but not the quadratic trend component ($p=.488$). Furthermore, surfaces also had a significant main effect and the interaction of surfaces with degrees of rotation was not significant ($p=.836$). Figure 5-8 presents the mean response times for degrees of rotation by the number of surfaces carried is additive, when summed over distractor sets.

Table 5-2 also shows a significant effect for the distractor set ($p<.005$). Further, however, the distractor type interacted significantly with the degrees of rotation. Figure 5-9 plots the relationship of mean

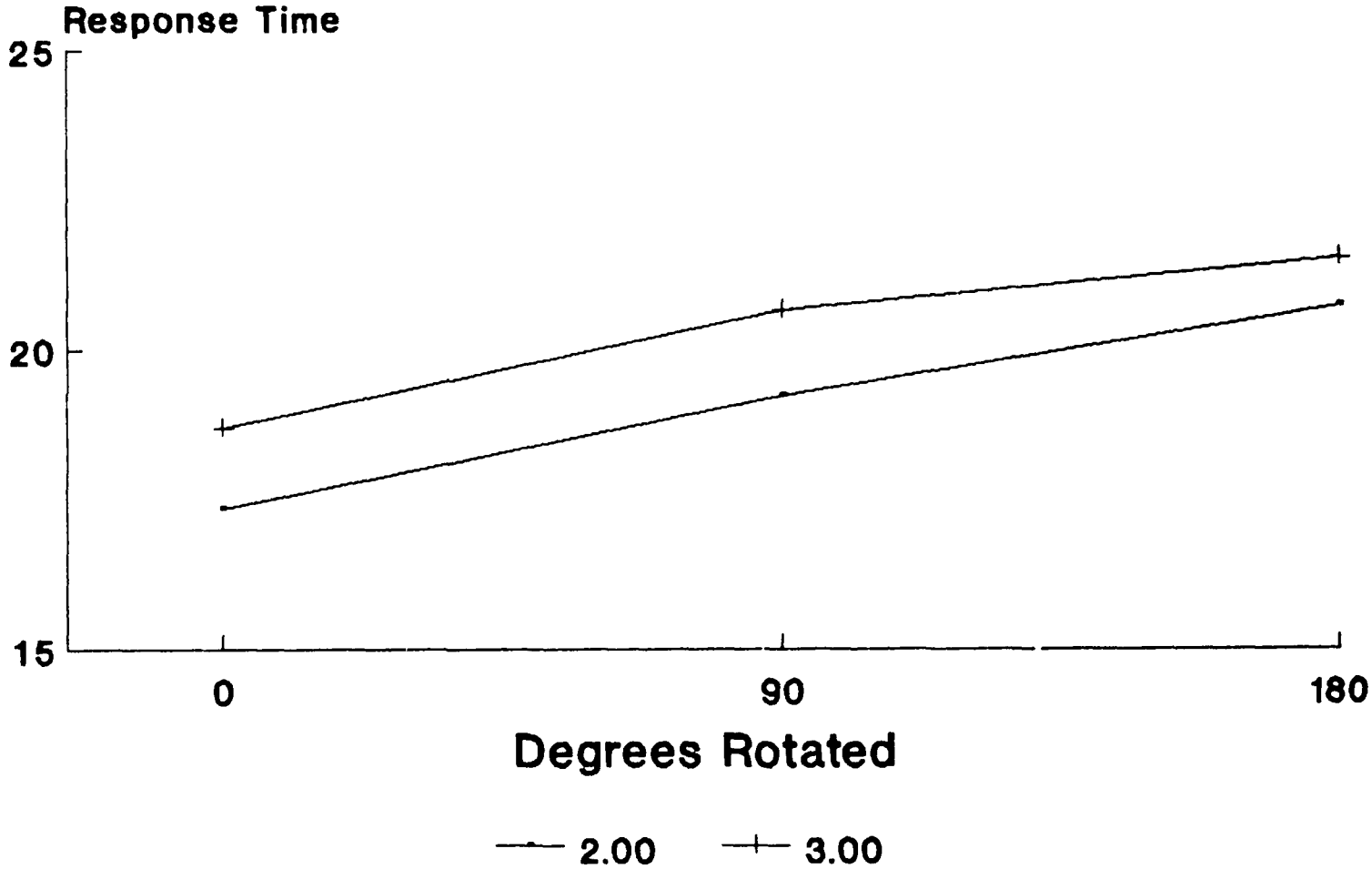
Table 5-2

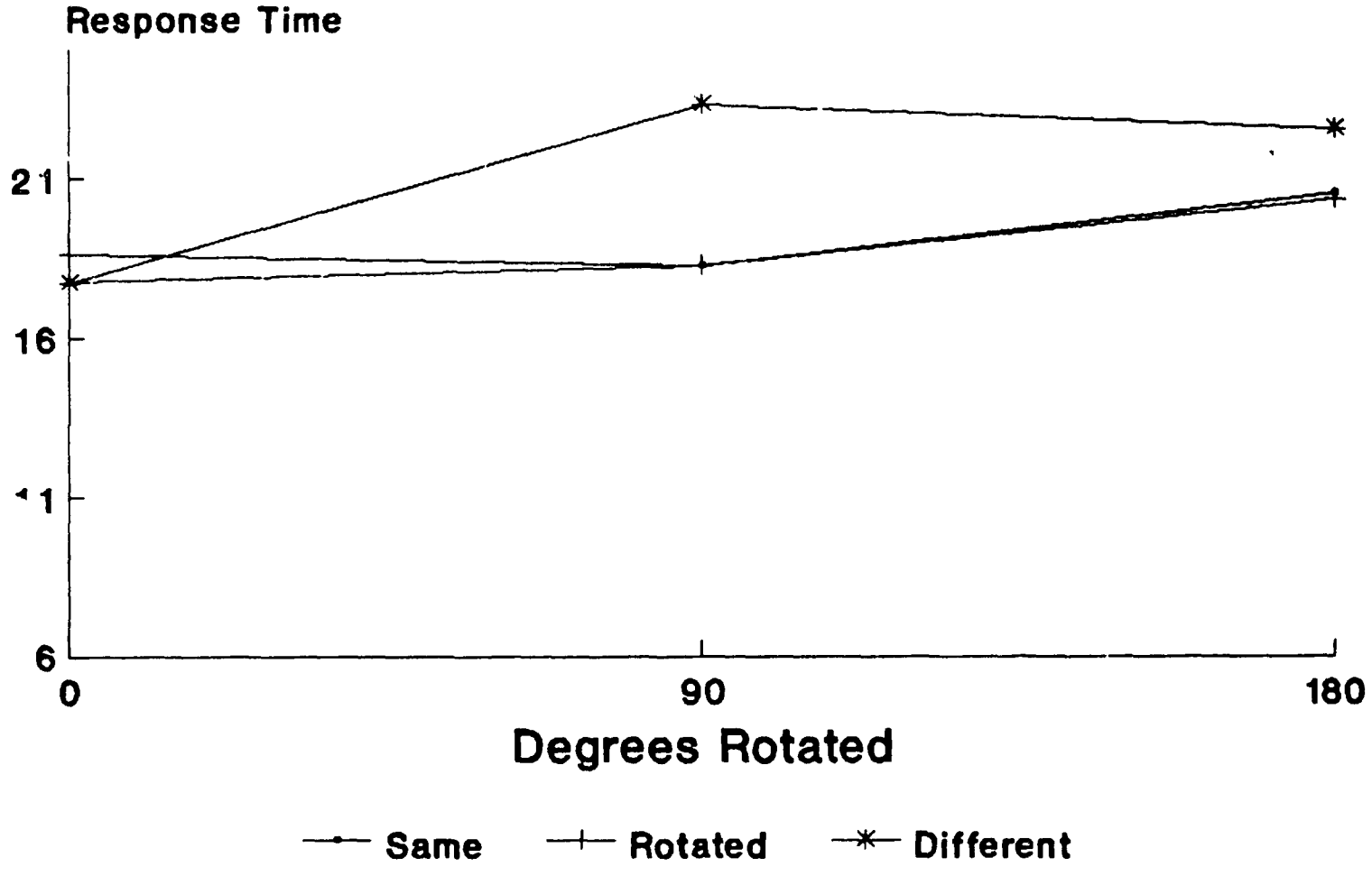
Within Subject Analysis of Correct Response Time for Spatial Items

Effect	df	MS	F	P
Degrees Rotation[1]	2	591.65	14.61	.000
Error	78	40.50		
Surfaces	1	257.45	4.38	.043
Error	39	58.45		
Distractor Type	2	400.23	5.64	.005
Error	78	70.97		
Degrees X Surfaces	2	7.12	.18	.836
Error	78	39.63		
Degrees X Distractor	4	208.61	5.26	.001
Error	156	39.63		
Surfaces X Distractor	2	11.21	.23	.793
Error	78	48.10		
Degrees X Surfaces X Distractors	4	25.71	.56	.695
Error	156	46.29		

[1] Linear Effect, $P < .0009$

Quad Effect, $P = .488$


Response Time
25
20
15
0
90
180
Degrees Rotated
2.00
3.00


Response Time
21
16
11
6
0
90
180
Degrees Rotated
Same
Rotated
Different

response times to degrees of rotation for each distractor set. It can be seen that processing durations increase only for the correct answers with 180 degrees of rotation within both the same and rotated distractor sets. However, for distractor sets with both different anchoring points and different orientations, processing durations are greater for both correct answers with 90 and 180 degrees of rotation.

Table 5-3 presents a within subjects analysis of variance on correct response time for the adjacent relation problems. The Degrees effect was highly significant. The degrees of rotation effect was linear, as the linear component was highly significant ($p=.003$), but not the quadratic component ($p=.009$). However, the interaction of Degrees and Distractors was also significant. The trend contrasts indicated that the interaction involved the quadratic component ($p=.009$), but not the linear component ($p=.421$). Figure 5-10 shows that the major difference between distractor sets occurred at 90 degrees of rotation. For the distractors with the same orientation, the problems with a 90 degree rotation of the correct answer were relatively easy, while for the distractors with rotated orientations, a 90 degree rotation of the correct answer was relatively harder.

Table 5-4 presents the results of a within subjects analysis of variance for response times on the position relation problems. Degrees of rotation had a significant effect, and the trend analysis showed that both the linear component ($p=.013$) and the quadratic component ($p=.016$) were highly significant. The Distractor effect was not significant, but Degrees X Distractor interaction was highly significant ($p<.001$). Figure 5-11 plots the mean response times for the position problems. It can be seen that for distractor sets with both different orientation and adjacent

Table 5-3

Within Subjects Analysis of Correct Response Time on Adjacent Relation Problems

Effect	df	MS	F	P
Degrees[1]	2	211.88	6.72	.002
Error	78	31.51		
Distractor Type	1	31.45	1.06	.310
Error	39	29.77		
Degrees X Distractor[2]	2	75.79	3.52	.034
Error	78	21.50		

[1] Linear P = .003

Quad P = .489

[2] Linear P = .421

Quad P = .009

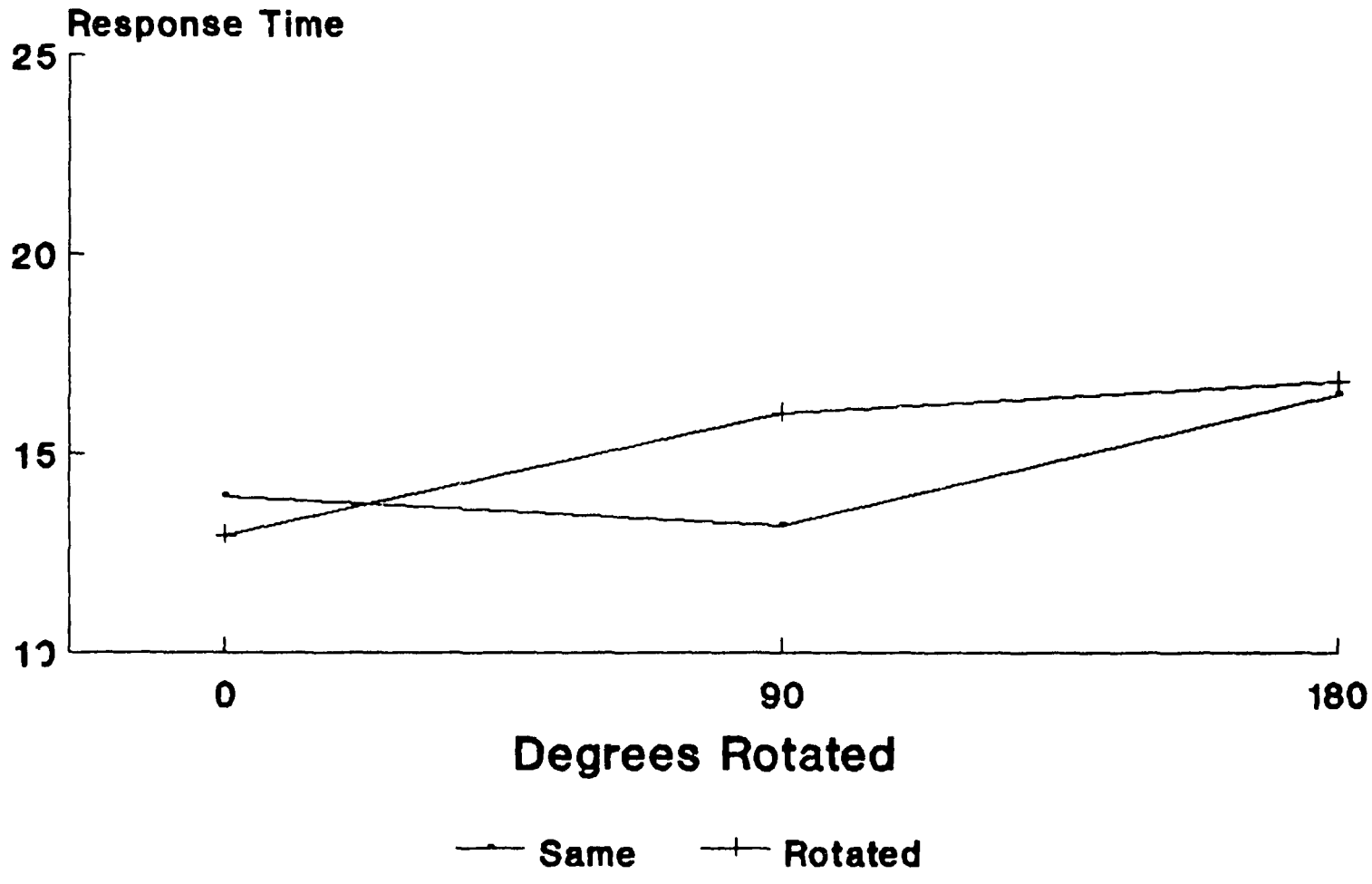

Response Time
25
20
15
10
0
90
180
Degrees Rotated
Same
Rotated

Table 5-4

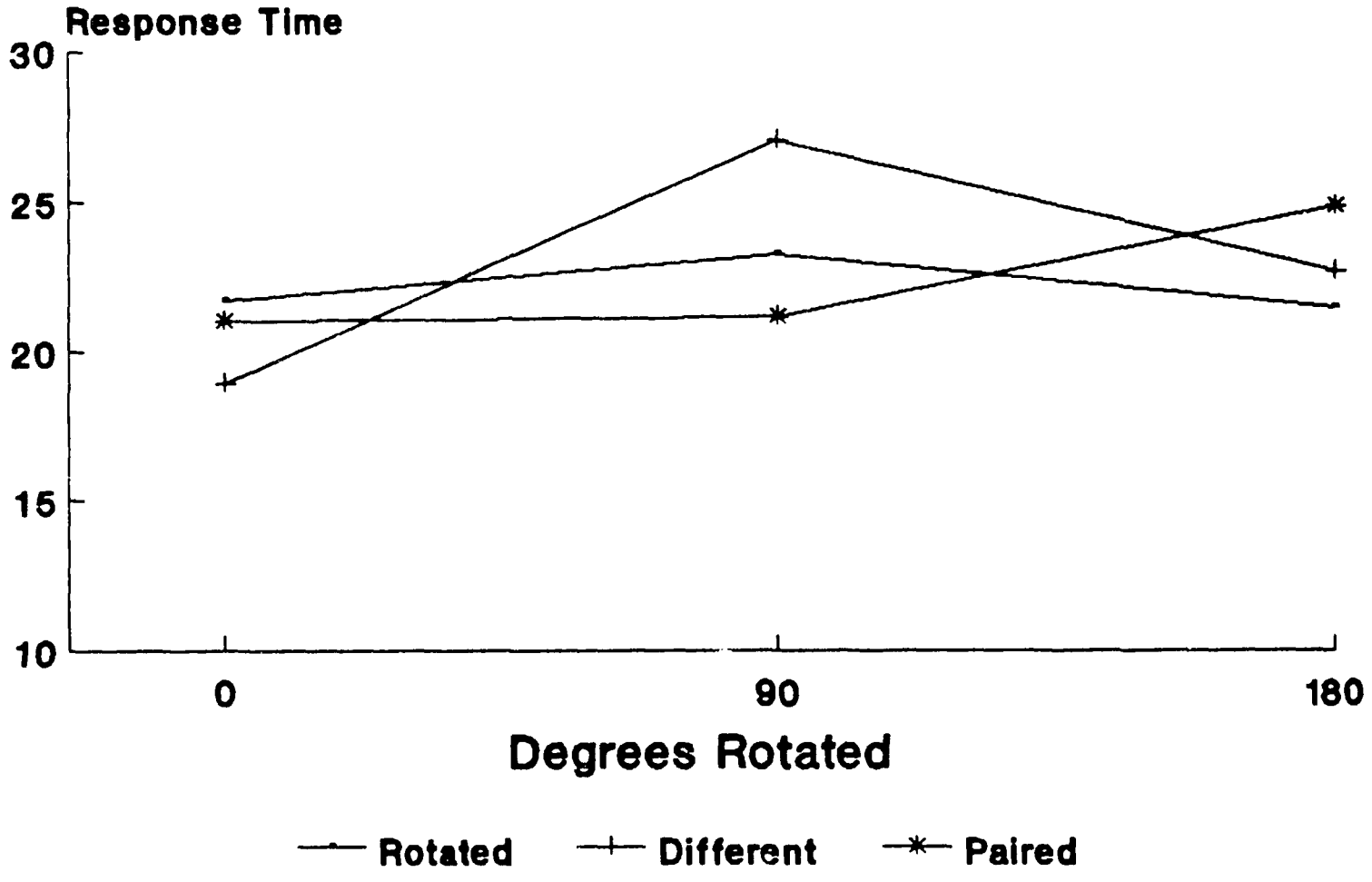
Within Subject Analysis of Variance on

Correct Response Time for Position Problems

Effect	df	MS	F	P
Degrees[1]	2	345.31	6.57	.002
Error	78	52.54		
Distractor Type	2	17.92	.33	.72
Error	78	54.13		
Degrees X Distractors	4	272.14	5.04	.001
Error	156	54.03		

[1] Linear Effect P = .013

Quad Effect P = .016


Response Time
30
25
20
15
10
0
90
180
Degrees Rotated
Rotated
Different
Paired

anchoring pairs, correct answers that were rotated 90 degrees had longer processing durations than either the correct answers with 0 or 180 degrees of rotation. In contrast, for the pair distractor sets, the correct answers with 180 degrees rotation had longer processing durations.

Accuracy. Figure 5-12 plots the mean accuracies by problem type and distractor sets, within problem types. Table 5-5 presents the corresponding within subject analysis of variance for accuracy, in which Distractors is nested within Problem Type. It can be seen that relatively higher accuracy of the adjacent relation problems, as compared to the spatial and positional problems, is highly significant ($p=.013$). Furthermore, the spatial problems had a significantly higher accuracy than the position relation problems ($p=.023$).

Within the spatial problems, accuracy was significantly harder when the problems had distractors with the same anchoring point (presented in either the same or rotated orientations) than when the problems had distractors with different anchoring points. However, problems in which the distractors had the same orientation did not differ in accuracy from problems in which the distractors had rotated orientations.

Within the adjacent relation problems, the distractor sets did not differ in accuracy. However, within the position relation problems, the distractor sets with same orientation and paired orientation problems were solved less accurately than the problems in which the distractors had different anchoring points. The difference between the problems with rotated and paired distractors did not reach significance.

The analyses of variance for the other effects were conducted separately within problem type. Figure 5-13 plots the mean accuracies for

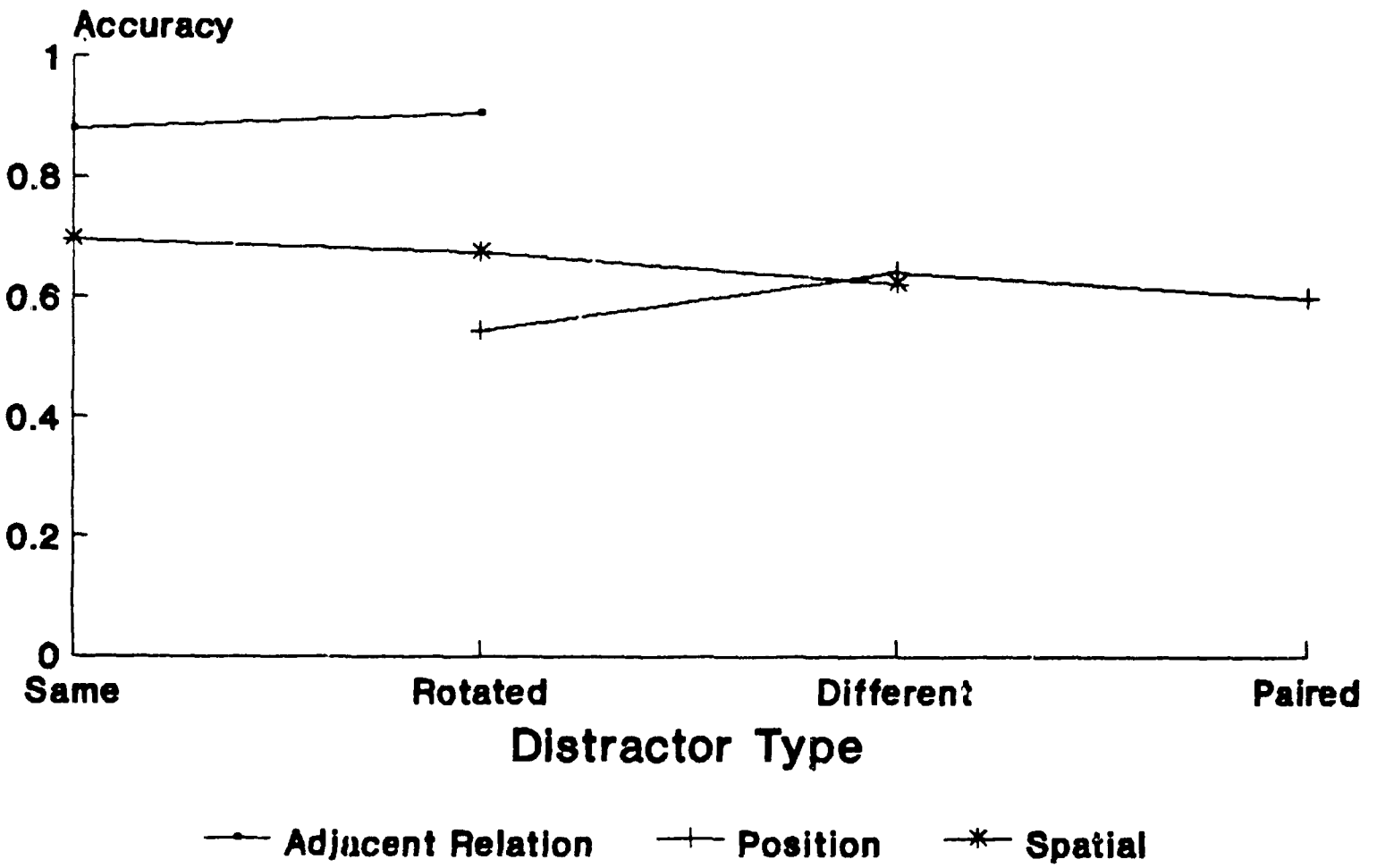
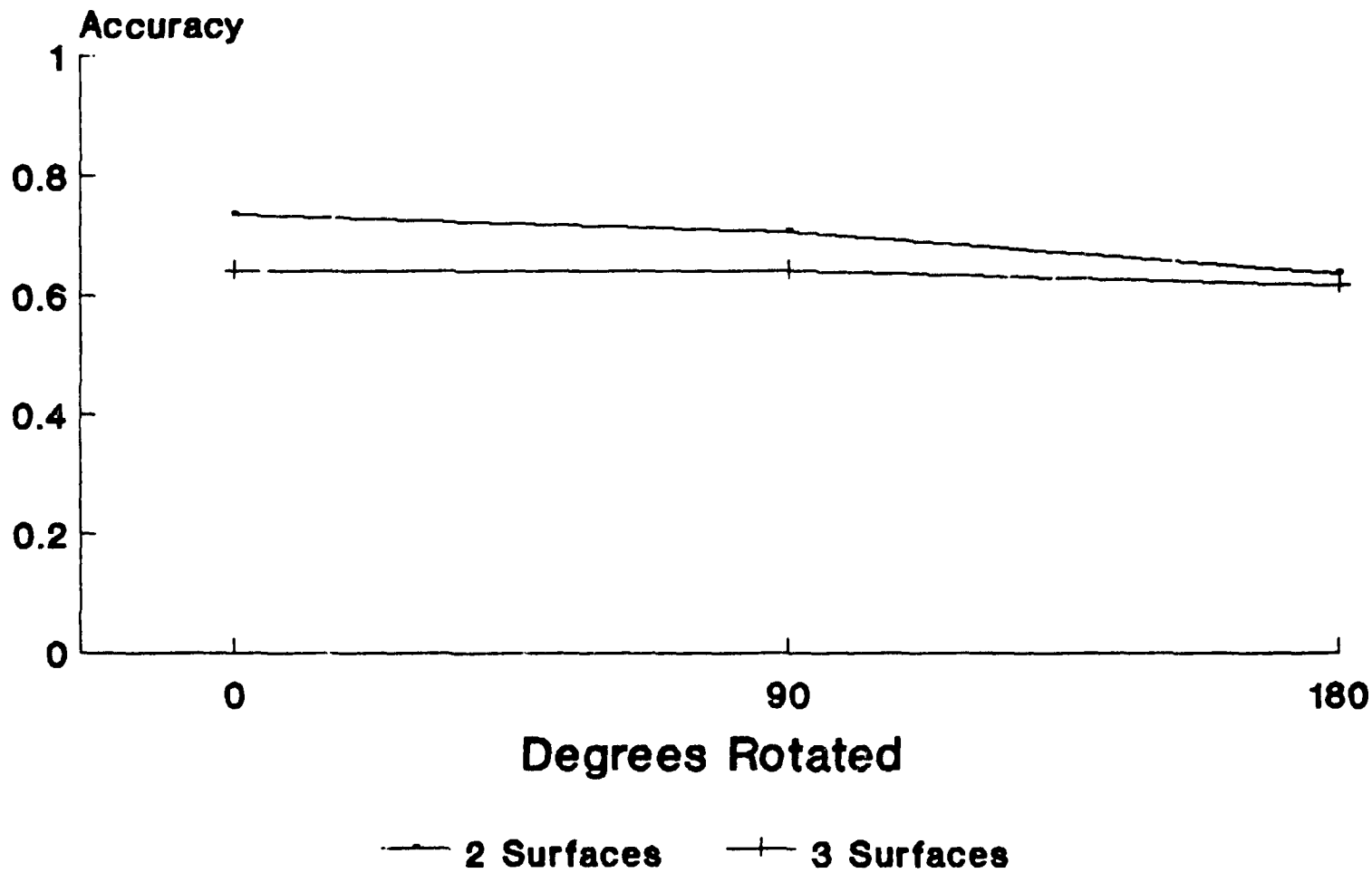

Accuracy
1
0.8
0.6
0.4
0.2
0
Same
Rotated
Different
Paired
Distractor Type
Adjacent Relation
Position
Spatial

Table 5-5

Within Subject Analysis of Variance for

Accuracy By Problem Type and Distractor Set

Source	MS	F	P
Problem Type			
Spatial and Position vs. Adjacent Relation	258.38	147.70	.000
Spatial vs. Position	17.76	5.57	.023
Distractors Within Problem Type			
Spatial			
Same and Rotated vs. Different	5.69	10.38	.002
Same vs. Rotated	.31	.92	.341
Adjacent Relation			
Same vs. Rotated	.80	1.03	.317
Position			
Rotated and Paired vs. Different	7.83	4.22	.047
Rotated vs. Paired	5.94	3.24	.079


Accuracy
1
0.8
0.6
0.4
0.2
0
0
90
180
Degrees Rotated
2 Surfaces
3 Surfaces

the three degrees of rotation by the number of surfaces carried for the correct answer, summed over distractor effects. Table 5-6 shows that the main effects for both degrees of rotation ($p=.013$) and surfaces carried ($p=.004$) are significant, but not the interaction ($p=.117$). Further, a trend analysis indicated that the effect for degrees of rotation was linear, as the linear component was highly significant ($p=.009$) but not the quadratic component ($p=.373$).

However, Table 5-6 shows that although distractor set had a significant main effect ($p=.001$), it also had a significant two way interaction with degrees of rotation and a significant three-way interaction with degrees of rotation and the number of surfaces carried. Figure 5-14, Figure 5-15 and Figure 5-16 plot mean accuracies for the three degrees of rotation by the number of surfaces carried, separately within each distractor set. The degrees of rotation and number of surfaces carried had apparent effects on accuracy for both the problem in which the distractors had the same anchoring point. However, the degrees of rotation and the number of surfaces carried had little apparent impact on the problems in which the distractor had different anchoring points.

Figure 5-17 plots the mean accuracies over degrees of rotation for the two distractor sets in the adjacent relation problems. Table 5-7 shows that neither the main effects or interactions are significant.

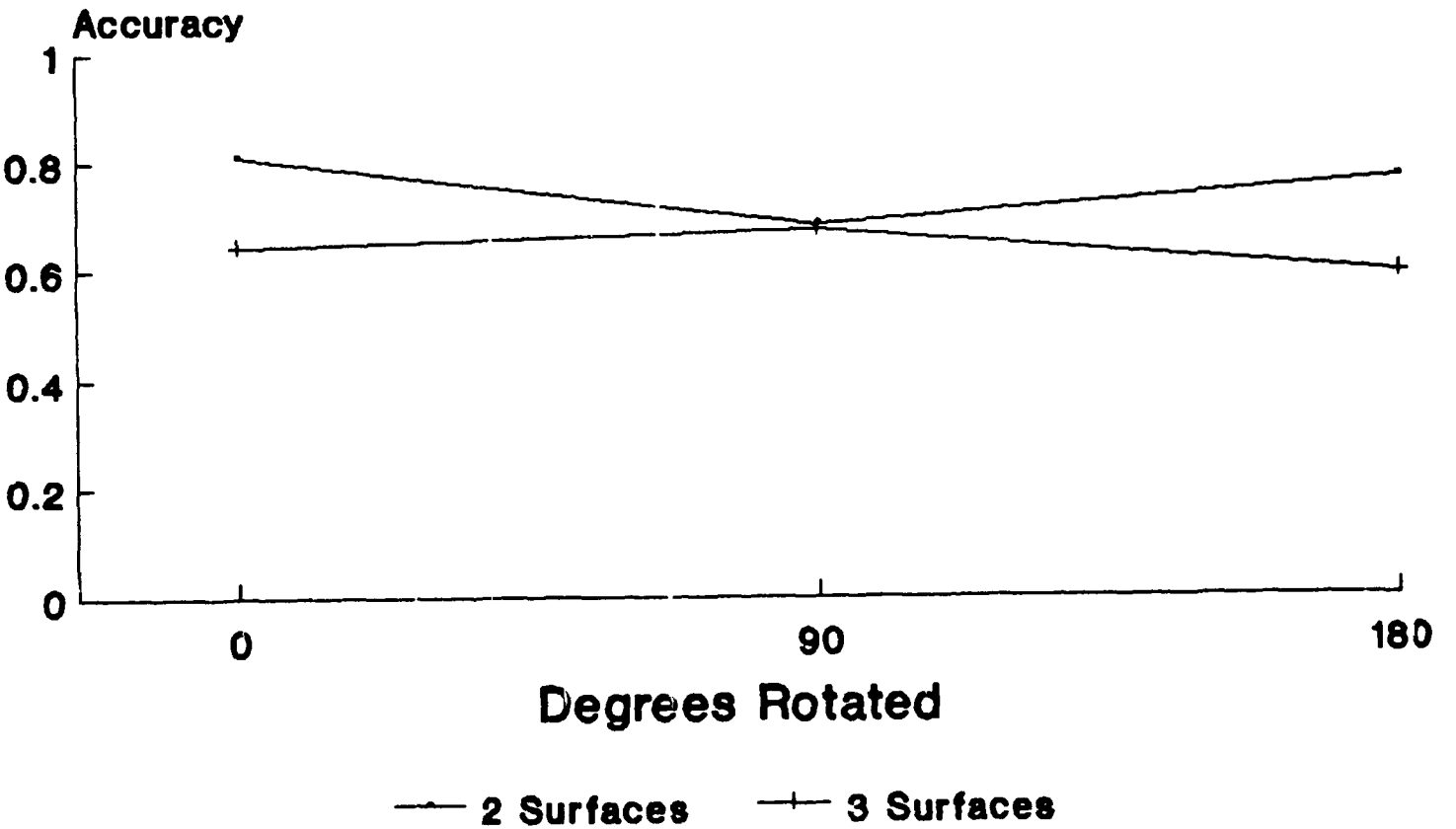
Figure 5-18 plots the mean accuracies over degrees of rotation for the distractor sets in the position problems. Table 5-8 shows that the main effect for degrees of rotation is highly significant ($p=.000$), as is the interaction of degrees of rotation with the distractor set ($p=.000$). A trend analysis revealed that degrees of rotation had both a significant

Table 5-6

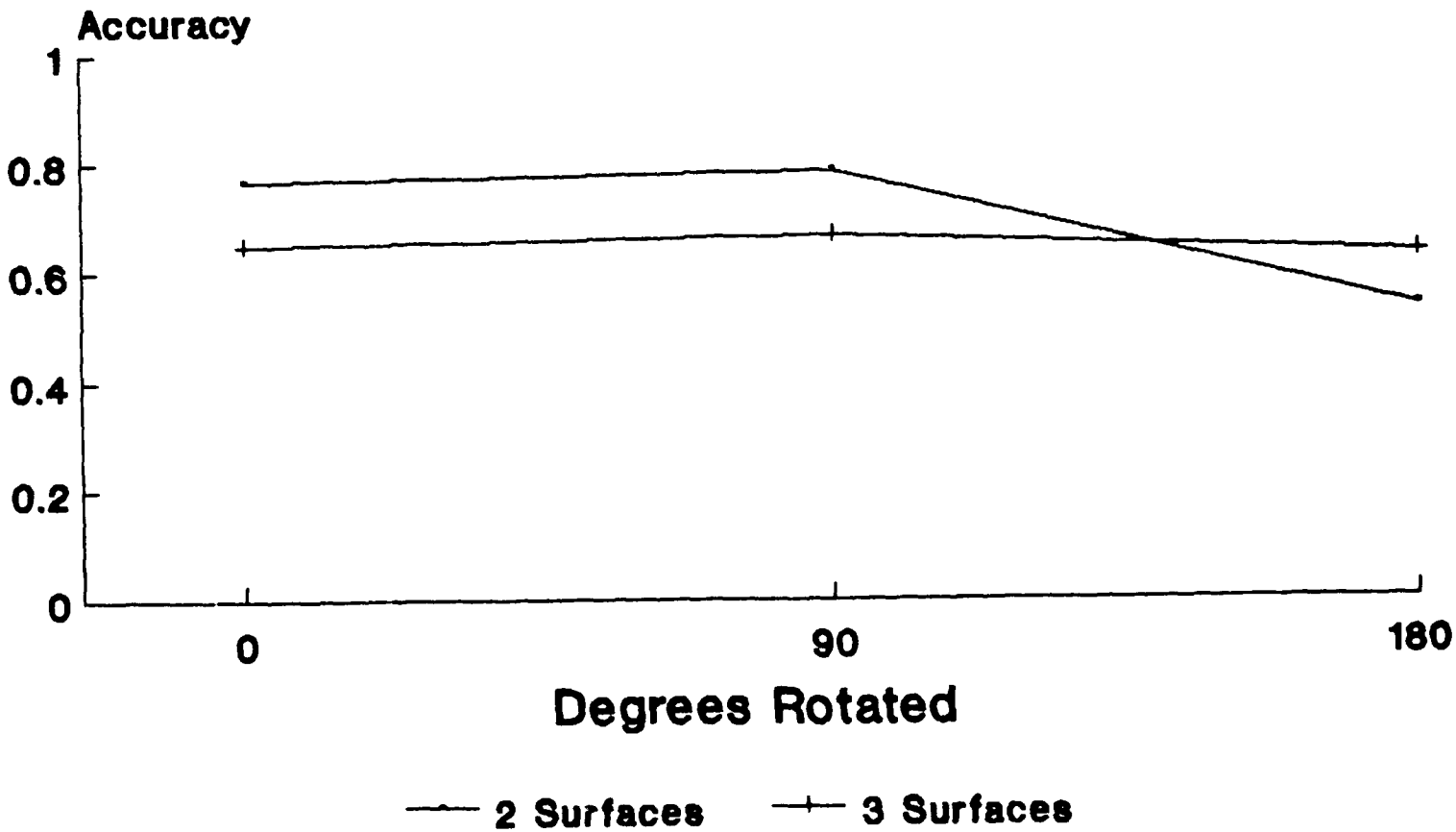
Within Subject Analysis of Variance on Accuracy for Spatial Problems

Effect	df	MS	F	P
Degrees	2	.25	4.56	.013
Error	78	.06		
Surfaces	1	.63	9.52	.004
Error	39	.07		
Distractor Type	2	.33	7.81	.001
Error	78	.04		
Degrees X Surfaces	2	.09	2.25	.117
Error	78	.04		
Degrees X Distractors	4	.14	2.35	.056
Error	156	.06		
Surfaces X Distractors	2	.16	2.05	.136
Error	78	.08		
Degrees X Surfaces X Distractors	4	.22	3.60	.008
Error	156	.06		

Same Distractors



Rotated Distractors



Different Distractors

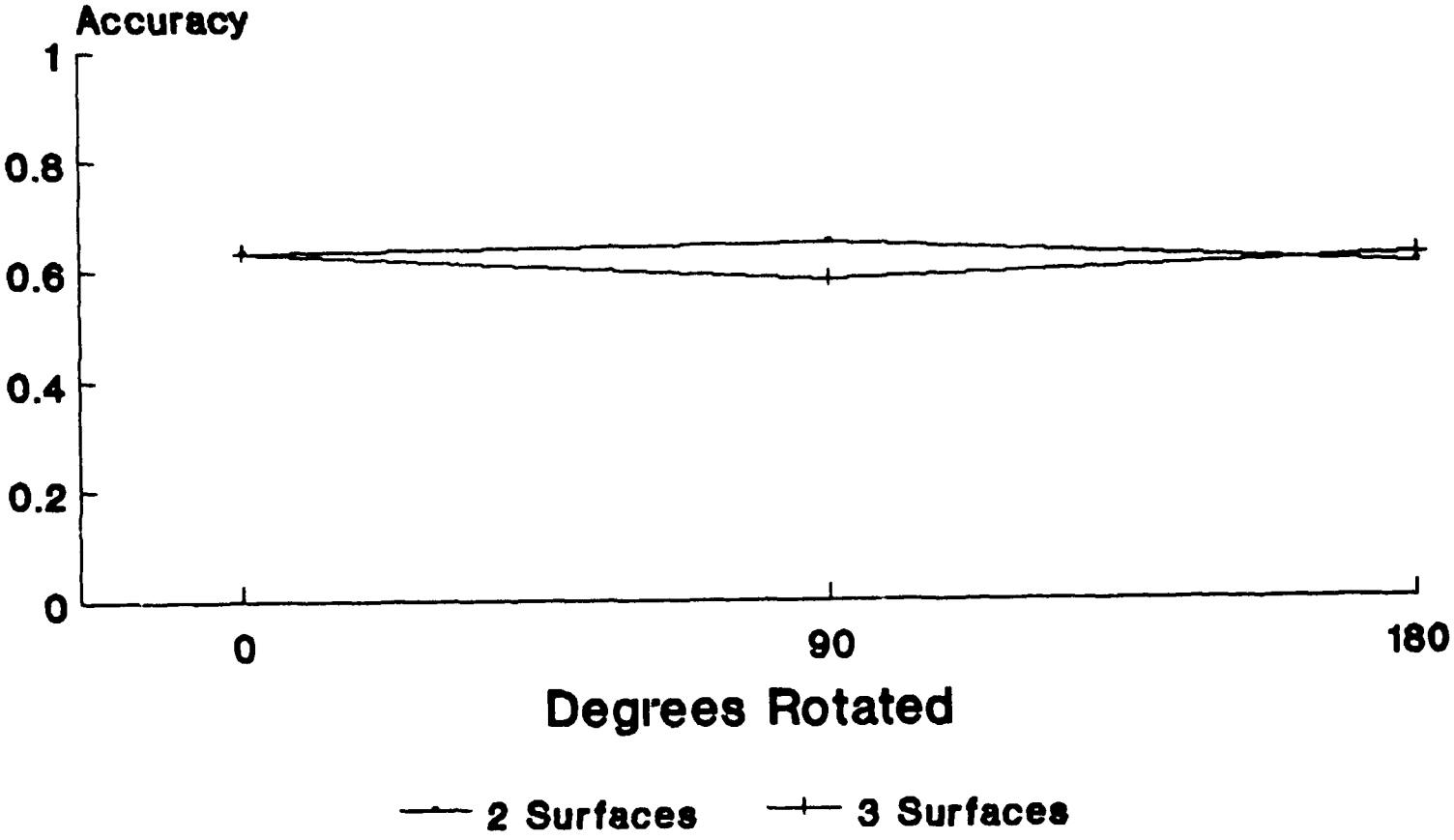


Table 5-7

Within Subject Analysis of Variance on Accuracy

for Adjacent Relation Problems

Effect	df	MS	F	P
Degrees	2	.00	.02	.979
Error	78	.02		
Distractor Type	1	.04	1.04	.314
Error	39	.04		
Degrees X Distractors	2	.07	1.93	.152
Error	78	.04		

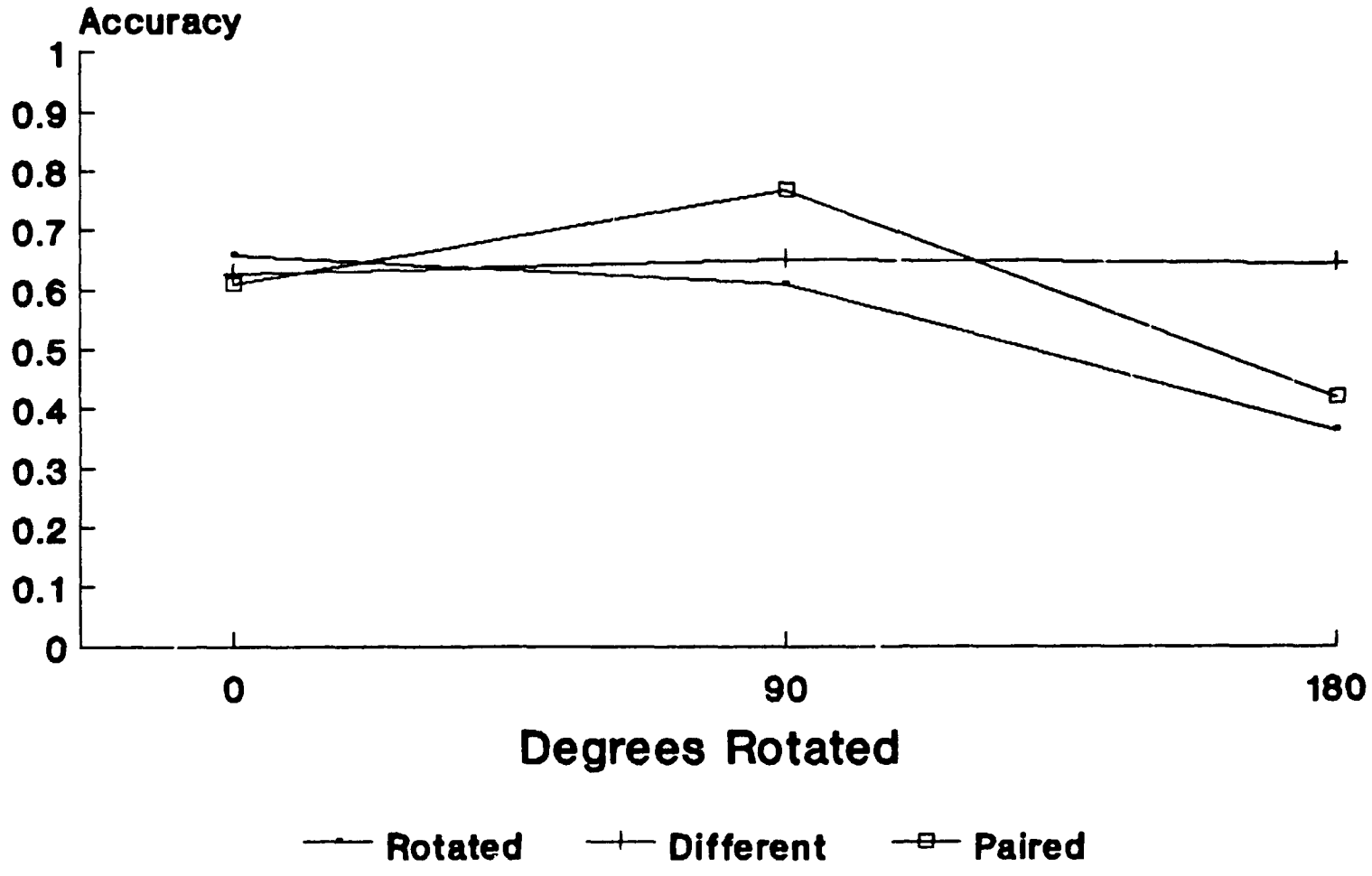

Accuracy
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
90
180
Degrees Rotated
Rotated
Different
Paired

Table 5-8

Within Subject Analysis of Variance of Accuracy

on Position Problems

Effect	df	MS	F	P
Degrees[1]	2	1.34	19.12	.000
Error	78	.07		
Distractor Type	2	.28	2.80	.067
Error	78	.10		
Degrees X Distractors	4	.45	6.07	.000
Error	156	.07		

[1] Linear $P < .0009$

Quad $P < .0009$

linear component (p=.000) and a quadratic component (p=.000). Figure D18 shows that the degrees of rotation for the correct answer had little influence on accuracy for problems in which the distractors had different anchoring points. However, the problems with 180 degrees of rotation for the correct answer had substantially lower accuracies for both problems with the rotated orientation and paired orientation distractors.

Mathematical modeling of response time and accuracy. The preceding analyses show the effects of the manipulations of the spatial folding task. However, they do not directly test the three alternative models, the attached folding model, the verbal-analytic model and the direct folding model, as these models require greater operationalization on the various stimuli. Extensive regression modeling for response time and accuracy are planned for this data.

Table 5-9 shows the results of a preliminary regression analysis that includes a partial operationalization of variables from each model. It can be seen that the models show moderately high prediction of both response time (R=.63) and accuracy (R=.67), which is measured as log odds (ln(P/1-P) so as to be scaled similarly to an item response response model (e.g., the Rasch model). Problem type (spatial and position problems versus adjacent relation problems) had the largest correlation with both response time and accuracy. Surfaces carried had the next largest correlation for both, followed by degrees of rotation. These results, then, support the major modeling variables in the study.

Additional models will contain independent variables which operationalize each processing model separately, so that their relative goodness of fits may be compared.

Table 5-9

Regression Modeling of Response Time and Accuracy[1]

	Response Time (R = .63)				Accuracy[1] (R = .67)			
Independent Variables	r	b	t	p	r	b	t	p
1) Spatial and Position versus Adjacent	.52	.10	2.83	.01	-.53	-.33	-2.84	.01
2) Spatial versus Position Problems	-.08	-.08	2.32	.02	.09	.27	2.48	.02
3) Degrees of Rotation	.23	.01	2.69	.01	-.24	-.01	-2.61	.01
4) Surfaces Carried	.41	.09	1.35	.18	-.42	-.34	-1.55	.12
5) Distractors within Spatial (Anchors)	-.18	-.05	1.96	.05	.12	.07	.91	.37
6) Distractors within Spatial (Orientation)	.00	.01	.13	.89	-.03	-.08	-.55	.58
7) Distractors within Adjacent Relation	.00	.01	.14	.89	.04	.10	.53	.60
8) Distractors within Position Anchors	.06	.03	.78	.44	-.08	-.16	-1.42	.16
9) Distractors within Position (Orientation)	-.06	-.04	-.76	.45	.11	.26	1.34	.18
10) Rotational Direction	.12	.01	.17	.86	-.25	-.37	-2.11	.04

[1]Accuracy is measured as ln (P/1-P)

Strategy self-reports. Table5-10 presents the correlations between the seven questions about the strategies used during the spatial folding test. Although the correlations are not high, a Bartlett's test for the matrix indicated that the correlations were significant ($p<.01$). A maximum likelihood factor analysis revealed that two factors were sufficient to account for the pattern of correlations. The varimax rotated factors are presented on Table 5-11. The variables that load highly on the first factor a direct folding strategy, while the variables that load highly on the second factor describe a verbal-analytic strategy. Interestingly, the items about the attached folding model had low communalities, and further were uncorrelated with each other.

Further analyses are planned to examine if the mathematical model for the various strategies had differential goodness of fit for subjects who report using different strategies.

Discussion

A major finding from this study is that the attached folding model is empirically plausible for multiple choice spatial folding items which are not readily solvable by verbal analytic processes. Degrees of rotation and the number of surfaces carried had additive effects for both response time and accuracy. Further, the contrasts for trend showed that degrees of freedom had a linear relationship to both response time and accuracy, as in other studies of spatial processing.

A further finding is that problems that can be solved by adjacent relations are processed more quickly and more accurately than the spatial problems. However, the problems in which position relations are sufficient for solution were slightly more difficult than the spatial problems. The

Table 5-10

Correlations Between

Self-Report Strategy Questions

Question	1	2	3	4	5	6	
1 Use adjacent relationships	1.00						
2 Rotate stem before folding	.21	1.00					
3 Fold stem before considering alternatives	-.04	-.05	1.00				
4 Fold stem and then mentally rotate to alternative	-.02	-.15	.35	1.00			
5 Attach, rotate and fold	.21	.18	-.01	-.29	1.00		
6 Rotate and fold	.25	.19	-.12	.01	.12	1.00	
7 Use positional relationships	.45	.14	-.03	.07	-.06	.05	1.00

Table 5-11

Factor Loading Matrix for

Strategy Questionnaire

Question	FI	FII	h^2
1 Use adjacent relationships	-.01	.40	.48
2 Rotate stem before folding	-.14	.11	.15
3 Fold stem before considering alternatives	.35	-.03	.13
4 Fold stem and then mentally rotate to alternative	.99	.06	.99
5 Attach, rotate and fold	-.27	-.09	.24
6 Rotate and fold	.04	.01	.19
7 Use positional relationships	.02	.99	.99

latter findings seems surprising, on the surface. However, the difficulty of the position relations problems probably results from the large proportion of spatial problems in the item set. Although the position relationship was sufficient, it would be largely unsuccessful in the item set as a whole, since distractors could not be falsified by this strategy for the spatial items. More stringent comparisons of cognitive models for the various item types are planned for future data analysis.

Another important finding was that the nature of the distractors influenced either the amount, difficulty and nature of processing in all but the adjacent relation problems. For the spatial problems, the distractor set characteristics had a significant main effect on response time and accuracy. Further, however, the distractor set characteristics interacted with the degrees of rotation required to confirm the correct answer. Thus, the nature of processing was also influenced by the distractors.

The major differences result from distractor sets in which the anchoring points vary versus the distractor sets in which the anchoring points were constant. It would appear that the characteristics of the correct answer had less impact on processing when new anchor points are required to falsify the distractors. Mathematically modeling, in which the independent variables more explicitly operationalize the differences between the three postulated information processing models, may elaborate this effect more completely.

In summary, these data show that the attached folding model is sufficiently supported to be a viable basis for constructing a spatial test in which the cognitive load is carefully specified. The attached folding

model is empirically plausible for multiple choice spatial folding items.

Calibration of the Spatial Learning Ability Test

Obtaining a large sample data for the calibration of Spatial Learning Ability Test (SLAT) was a major goal for the project. The plan specified calibrating spatial learning ability from the multidimensional item response model for learning and change (Embretson, 1987; 1989a; 1989b), which requires large sample sizes. The facilities at Lacklund Air Force base provided an ideal opportunity for collecting calibration data, as it has a large microcomputer laboratory and subjects who are available for a relatively long testing session.

The Spatial Learning Ability Test is described in detail in the methods section. The specific procedures that were applied to the Lacklund sample are also described in the methods section.

Method

Test design: Cognitive features. Figure 5-1 shows an item that from the spatial folding item bank. All items contain a stem, which can be folded into a cube, and four folded alternatives. Subjects are instructed to fold the stem down mentally, to make a cube. Each side of the cube has a directed marking (e.g., a black arrow) that appears different when the cube is rotated into various positions. The correct answer is a three-dimensional view of three adjacent sides of the folded stem. The distractors are views of a different folded stem, in which the markings are in different relative positions.

The test battery consists of three counterbalanced testing forms, Form M1, Form M2, and Form M3, four linking items and two training units with practice items. These materials are described in detail here.

The test forms, Form M1, Form M2 and Form M3, contain 24 items each,

in which items are counterbalanced to represent stimulus design features that influence cognitive load. Eighteen items in each form represent the variables of the attached folding model, which includes the degrees of rotation to correct answer (0, 90, 180),the number of surfaces carried (1,2,3) and the distractor type (Same or Rotated). For the one-surface items, a verbal-analytic strategy also could be applied to solve the task. The remaining six items in each form could be solved by another verbal-analytic strategy, using position relations. These items vary in the distractor type (Rotated or Paired) and in the degrees of rotation (0, 90, 180). All position relationship items were two surface problems.

Three different stem shapes appear on the items, the "T", "Z" and "F" shapes on Figure 3-1. Items were constructed to represent each combination of the design features for each stem shape. The design features appear equally on the three forms, but the exact stem shape that represents the combination varies between forms. Table 6-1 shows the Latin square design that assigns the three spatial folding shapes to the various design features of the attached folding model in each form, while Table 6-2 shows the design for the position relations items.

The order of the design features was randomly assigned. However, the various combinations of design features appear in the same order position in each form.

Intervention Conditions. In the first intervention condition, Analogue Training, the physical analogues of the mental folding task are presented to the subject.

That is, three wooden cut-outs that can be physically folded into a cube are given as instructional aids. The wooden cut-outs match the item stems

Table 6-1

Latin Squares to Assign Stem Shape to the Stimulus Design Variables in Three Forms

		Distractors-Same			Distractors-Rotated		
		Surfaces Carried			Surfaces Carried		
Degrees Rotation		1	2	3	1	2	3
Form 1							
	0	1	2	3	2	1	3
	90	3	1	2	1	3	2
	180	2	3	1	3	2	1
Form 2							
	0	3	1	2	3	2	1
	90	2	3	1	2	1	3
	180	1	2	3	1	3	2
Form 3							
	0	2	3	1	1	3	2
	90	1	2	3	3	2	1
	180	3	1	2	2	1	3

* Shape Codes

1 = 2 =

3 =

Table 6-2

Latin Squares to Assign Positional Items to Three Forms

Degrees of Rotation	Distractors: Same	Distractors: Rotated
Form 1		
0	1	3
90	2	1
180	3	2
Form 2		
0	2	1
90	3	2
180	1	3
Form 3		
0	3	2
90	1	3
180	2	1

* Shape Codes

1 - 2 -

3 -

that are shown on the computer screen. The instructions indicate that the orientation and position of the markings should be carefully observed while physically folding the wooden cut-out into a cube.

Five practice items are presented for each stem shape. The items vary in the degrees of rotation, the number of surfaces carried and the distractor types, as on the test items. No items that are presented with the Analogue Training can be solved by a simple verbal-analytic strategy. Thus, no one-surface problems or position relationship problems are presented.

In the second intervention, Strategy Training, the instructions concern the attached folding model and the verbal-analytic model for solving items. For the attached folding model, diagrams are presented with some items to show the process of locating adjacent sides, rotating to match the markers on the folded view and then folding to obtain the third side. The instructions note that the attached folding model can be applied to any item in the set. Several practice items are administered both with the instruction and following the instruction.

For the verbal-analytic model, the instructions indicate that the strategy can be applied only to some items. The instructions illustrate solving items by eliminating the distractors in which the markings on adjacent sides did not have the same relative position as in the unfolded stem view. The instructions include some simple position relations for the third side as another verbal-analytic strategy that could eliminate distractors on some items.

Subjects. The subjects were 582 Air Force recruits who were completing basic training at Lacklund Air Force Base. The recruits had

Table 6-3

Latin Square Design for Test Forms

by Testing Occasion and Groups

	Testing Occasion		
Group	Pretest	Posttest 1	Posttest 2
1	Form 1	Form 2	Form 3
2	Form 3	Form 1	Form 2
3	Form 2	Form 3	Form 1

completed approximately three weeks of basic training at the time of the testing session.

Procedure. Thirty microcomputers were used to administer the complete Spatial Learning Ability Test. The test was presented under three conditions of order for the test forms, as shown on Table 6-3. Thus, 10 microcomputers were devoted to each order. Recruits were randomly assigned to a microcomputer by their drill sergeant before entering the testing laboratory. The subjects were tested for approximately three hours.

The Spatial Learning Ability Test was administered prior to two other tasks. The other tasks consisted of the Mathematical Learning Ability Test and a training criterion task, the Logi-Gates electronics trouble-shooting task. The complete session lasted three hours. About two-thirds of the sample finished the first half of the Logi-Gates training.

Results

Descriptive statistics. Table 6-4 presents descriptive statistics for the raw test scores. It can be seen that accuracy levels, measured as log odds ($\ln(P/1-P)$) to be comparable to an item response model scale, increased from the pretest to the posttests. Further, mean response times decreased. Thus, performance was improving, as accuracy increased and processing time decreased.

Interestingly, the psychometric properties of the tests changed. The internal consistencies increased, and the variances and correlations between successive tests had a pattern that would be predicted by simplex model of performance trials. That is, the variances increased, and the correlations between successive tests increased as well.

Item characteristics and score distributions. Figure 6-1 presents a

Table 6-4

Descriptive Statistics for Spatial Learning Ability Test of Lacklund Sample

	Correlations			Cronbach's	Response Time		Log Odds Accuracy	
	(1)	(2)	(3)	Alpha	$\bar{X}$	SD	$\bar{X}$	SD
1) Pretest	1.00			.83	24.04	7.49	.12	1.14
2) Posttest 1	.79	1.00		.89	19.56	6.97	.65	1.47
3) Posttest 2	.76	.84	1.00	.90	15.54	6.10	.94	1.60

distribution of item accuracies on the pretest over the three groups. An inspection of Figure 6-1 suggests that the test is appropriate for the sample, as accuracy clusters at .50. Furthermore, the test also provides information at more extreme score levels, as some items with extreme accuracies are also included on the tests.

Estimating the parameters of the multidimensional latent trait model for learning and change requires the marginal frequencies of passing each item, as it appears on either the pretest, posttest1 or posttest2, within each group. Further, the marginal frequencies of total scores for each group on each test (pretest, posttest1 and posttest2) are also required. These marginal frequencies were obtained and program LINLOG was used to estimate the item parameters for the multidimensional latent trait model for learning and change. Table 6-5 shows the estimated item parameters.

Estimating the person's initial ability and two learning abilities, Analogue Learning Ability and Strategy Learning Ability, involves finding the ability estimates that maximize the likelihood of the person's responses over all conditions, given the item parameters. A multivariate Newton-Raphson procedure (Program MULTIDIM) was used to estimate the three abilities for each person. Distributions of abilities was then be prepared.

In addition to parameter calibration, the goodness of fit for the multidimensional item response model for learning and change was also assessed. Since the multidimensional model is a member of the Rasch family and has available conditional maximum likelihood estimates of the parameters (see Embretson, 1989), it will be possible to employ likelihood ratio significance tests similar to Andersen (1973) to assess model fit.

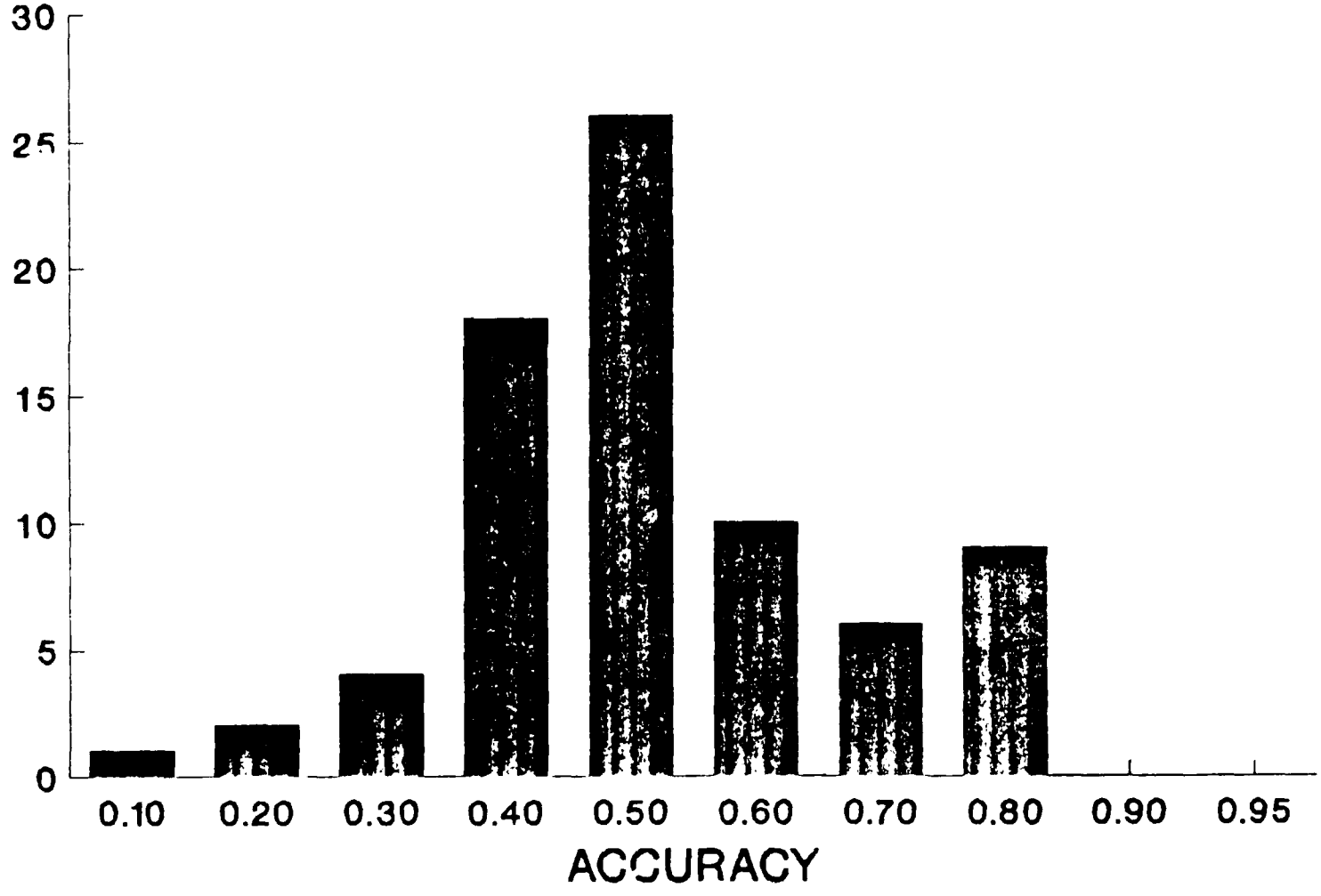

30
25
20
15
10
5
0
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
0.95
ACCURACY

Table 6-5

Spatial Item Difficulties from the Multidimensional

Latent Trait Model for Learning and Change

Form	Item	Sigma	Form	Item	Sigma	Form	Item	Sigma
M1	811	-.65328	M2	511	-.55290	M3	211	-.74534
	904	.68896		604	1.62345		304	1.36688
	308	-.45578		608	.12321		908	-.20143
	302	-.02644		602	.58775		902	.09228
	702	-.07361		102	.49311		402	-.08771
	link1	.16327		208	.30402		508	-.39222
	808	-.64326		606	1.60138		906	1.15397
	306	1.11607		509	-.46674		809	-.43070
	209	-.91105		108	.51202		408	-.37303
	708	-.35903		910	-.63015		310	-1.43928
	610	-1.23387		101	.73043		701	.25422
	link2	.26067		409	-.87494		709	-1.42693
	401	.39884		707	-.63015		107	-1.15750
	109	-1.53790		802	1.37605		202	.52534
	407	-.20613		804	.98180		504	.92961
	502	.65952		909	.74956		309	.13978
	204	.56215		710	-.10963		410	-1.59275
	link3	-1.31812		311	-.18899		911	-.73527

Table 6-5 (continued)

609	.06777	801	.39861	501	-.04037
110	-1.19889	104	.63518	704	.51553
611	-.62325	105	.72088	405	.15880
201	.08660	601	.73043	301	.01644
404	.42749	806	.60671	206	.22554
link 4	-.36539	210	.03666	510	-.51784
705	.32273				
901	.03951				
506	.24695				
810	-.43634				

Discussion

The analyses of the test calibration data from the Lacklund Air Force Base indicate that good psychometric properties have been achieved for the Spatial Learning Ability Test. Test variances, intercorrelations and internal consistencies had the hypothesized patterns. The item parameters have been calibrated and these parameters are sufficient to estimate the person parameters and prepare the norms.

Construct Validity for Measuring Spatial Learning Ability

A previous study (Embretson, 1987) had strongly supported the feasibility of measuring learning ability from the spatial folding task by using a physical analogue training procedure. The training procedure consisted of having subjects carry out in concrete the mental operations that the spatial task required, namely the folding and rotation of three dimensional shapes. Subjects were provided with a large physical replica of each stem, with prefolded edges, which they could manipulate to determine the (folded) alternatives that matched the drawing of the unfolded item stem. This intervention phase lasted only fifteen minutes, but resulted in an improvement of subjects' performances, not only during intervention but also afterwards, when the wooden cube was removed and subjects had to carry out the operations mentally again.

The results indicate that both aspects of construct validity (Embretson, 1983), construct representation and nomothetic span, were influenced by the dynamic testing procedure. Construct representation concerns the nature of the underlying processes, strategies and knowledge structures that are involved in performance. The following findings in Embretson (1987) indicate a change in construct representation: 1) the effect size for the increase in accuracy was about two-thirds of a standard deviation, 2) the analogue training results could not be attributed to a practice effect, and 3) the mental model that underlies performance was changed by the analogue training. For the latter, mathematical modeling of response accuracy indicated that a task characteristic that was postulated to influence the anchoring and confirmation process, was no longer significantly related to item

difficulty after the analogue training. Thus, these results indicate that the analogue training changed the construct representation of the spatial folding task.

Several findings also indicate that the nomothetic span of test scores was also changed. Nomothetic span concerns the properties of the test scores as a measure of individual differences. That is, the results indicate that 1) the internal consistency of the spatial test increase after analogue training, 2) analogue learning is person-specific, as indicated by LLTM model fitting, using Spada and McGaw's (1985) models, and 3) the learning ability measurements increase the predictive validity of the test for a technical training criterion. For the latter, the learning ability score had incremental validity for predicting vocational training in microcomputer operations over the initial spatial performance alone. These results support the psychometric utility of spatial learning ability.

In this section, two studies are presented on the construct validity of the Spatial Learning Ability Test. The first study is an analysis of relevant data from the Lacklund Air Force Base sample, while the second study is an experiment on a sample of college undergraduates. Both studies are partially analyzed at the date of this report.

Lacklund Study

The Lacklund calibration study also included data that is relevant to construct representation and nomothetic span. The relevant results include the impact of training on 1) performance levels, 2) performance dimensionality, 3) mental models that underlie performance and 4) predictive validity.

Method

The materials, procedures and sample were described in the preceding sections.

Results

Impact of training on performance. Figure 7-1 presents the accuracy levels at the pretest, first posttest and second posttest. It can be seen that accuracy increases sharply. The effect size for the mean change from the pretest to the first posttest, using the standard deviation from the pretest, is .47, which is similar to results obtained by Embretson (1987). The effect size from the pretest to the second posttest is .72. Thus, performance levels increased substantially.

The effects for both analogue training and strategy training on performance accuracy are highly significant. Table 7-1 presents an analysis of variance for the three testing occasions, over the three groups who received the test forms in a counterbalanced order. It can be seen that the effect of testing occasion was highly significant. Further, planned contrasts indicated that the change was significant both from the pretest to the first posttest ($p<.001$) and from the first posttest to the second posttest ($p<.001$), which supports the impact of analogue training and strategy training conditions, respectively, on accuracy.

Dimensionality of spatial ability. Another significant issue is the dimensionality of (effective) spatial ability over the successive occasions of measurement. In particular, the ability scores are expected to have certain psychometric properties if, in fact, learning abilities have been measured. In the multidimensional latent trait model for learning and change, it is postulated that learning abilities have additive

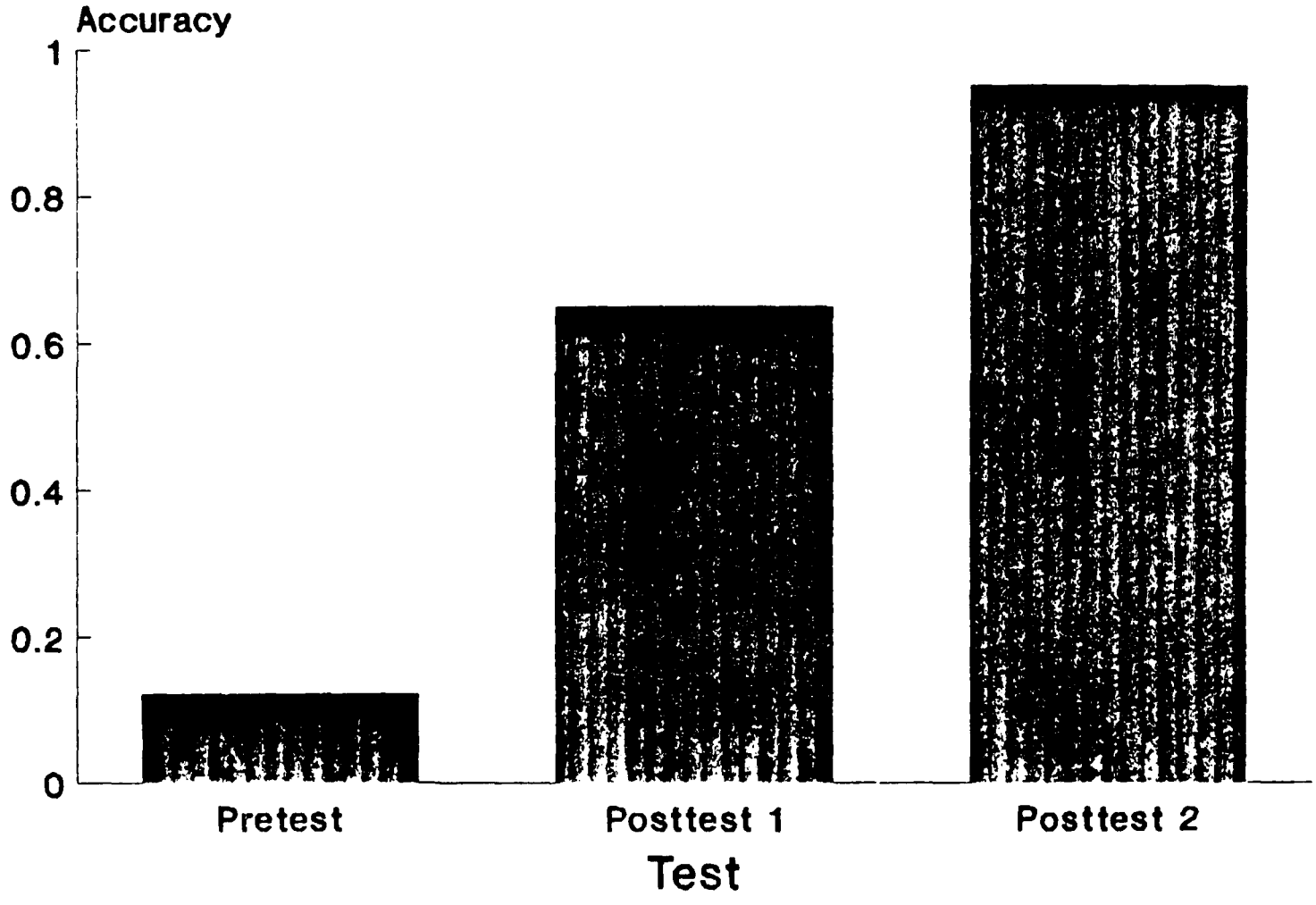

Accuracy
1
0.8
0.6
0.4
0.2
0
Pretest
Posttest 1
Posttest 2
Test

Table 7-1

Analysis of Variance for Spatial Learning Ability on Lackland Sample for Log Odds Accuracy

Source	df	MS	F	P
Between Groups	2	6.05	1.19	.31
Error	579	5.10		
Within Groups				
Test	2	99.85	224.95	.000
Test X Group	4	.32	.73	.572
Error	1158	.44		

effects on latent response potential. If the model is true, then increased internal consistencies, test correlations and test variances are expected over the successive ability scores. That is, when abilities are additive and response errors are constant, test variance is expected to increase over testing occasions, due to the absolute increase of true variance. Also, under these conditions the relative proportion of true variance increases, so that both increasing internal consistencies and increasing test correlations are also expected. The results (presented previously on Table 6-4) showed the expected pattern.

Impact of training on mental models. The general impact of the two training conditions on the mental models underlying performance is suggested by a change in processing durations. Figure 7-2 presents the mean item response times for the pretest, first posttest and second posttest. It can be seen that response times decreased over the testing occasions. Table 7-2 shows that these changes were highly significant.

The more specific impact of strategy training on mental models can be obtained by comparing mathematical models of cognitive processes between the pretest, the first posttest and the second posttest. The variables of the mathematical model are the variables in the three cognitive models, the attached folding model, the direct folding model and the verbal-analytic model, as described above. Significant changes would be indicated by finding different weights for the variables that operationalize the models. In the attached folding model, for example, these variables include the degrees of rotation and the number of surfaces carried. These analyses will be completed after this report.

Predictive validity. A computerized training sequence for electronics

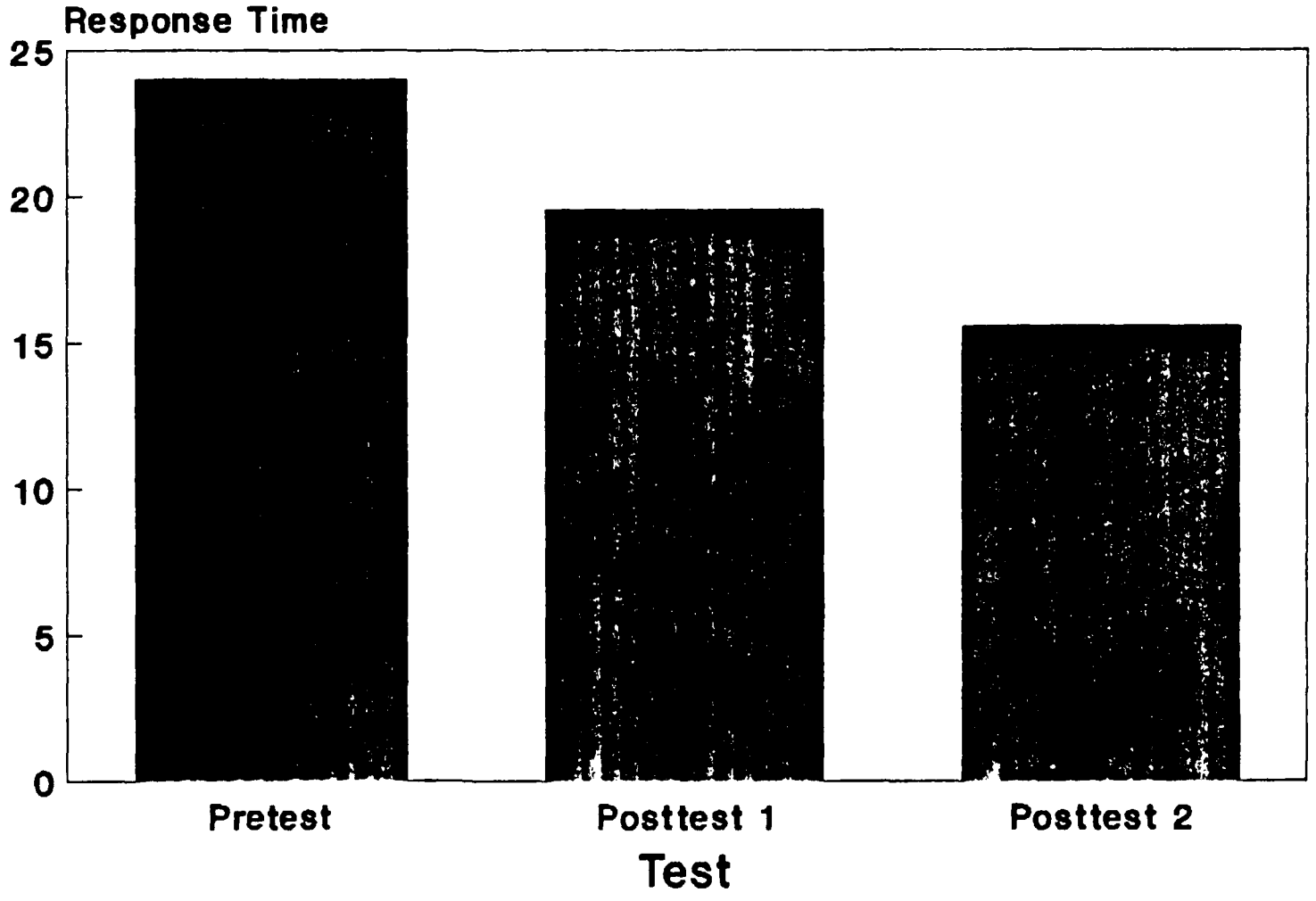

Response Time
25
20
15
10
5
0
Pretest
Posttest 1
Posttest 2
Test

Table 7-2

Analysis of Variance for Spatial Learning Ability on Lackland Sample for Response Time

Source	df	MS	F	P
Between Groups	2	108.58	1.02	.36
Error	577	106.29		
Within Groups				
Test	2	10486.95	595.80	.00
Test X Group	4	8.46	.48	.85
Error	1154	17.60		

Table 7-3

Descriptive Statistics for Training Success

on the Logic-Gates Electronic Tasks

		Accuracy		Response Time	
Trial	N	Mean	Standard Deviation	Mean	Standard Deviation
Without Negation					
Block 1	436	.783	.125	1.775	.577
Block 2	427	.853	.121	1.564	.490
Block 3	409	.893	.100	1.426	.425
Block 4	403	.917	.083	1.290	.368
Block 5	394	.932	.077	1.194	.309
Block 6	388	.936	.071	1.168	.310
Block 7	380	.949	.087	1.126	.317
With Negation					
Block 1	334	.751	.192	3.667	1.032
Block 2	302	.799	.181	2.784	.962
Block 3	268	.826	.167	2.542	.802
Block 4	242	.850	.154	2.540	.825
Block 5	224	.849	.168	2.523	.850
Block 6	207	.854	.156	2.375	.791

trouble-shooting, the Logic-Gates, was also administered to the Lacklund sample. The results from this task provide a criterion measurement to examine the predictive validity of the spatial learning abilities.

Table 7-3 presents the means and standard deviations for accuracy rates and response times over successive trials of the Logi-Gates electronics training task. It can be seen that accuracy increased and response time decreased over training trials for both types of Logic-Gates' tasks (with or without negation). Individual differences in accuracy rates, response times and efficiency rates (response time per accurate detection of electronics bugs) will be measured at the various trial blocks. The final trial block within each problem type will provide a learning criterion for which to examine the incremental validity of the spatial learning abilities when added to initial spatial ability. The validity analysis will also be repeated on residualized gain to the final trial, using the initial trial score as a covariate.

Discussion

The results obtained thus far support the test as measuring spatial learning ability. First, the spatial training had clear and substantial effects on performance levels. Second, the correlations and variances of test scores had the expected pattern for the measurement of learning ability. Third, a general impact of training on mental models was observed. More specific analyses are planned to determine the locus of this effect.

One remaining source of data from the Lacklund study, the criterion learning measurement, has not yet been fully analyzed. The criterion learning measurements will be regressed on the initial and learning

abilities to determine predictive validity. The initial analyses of the task, the Logic-Gates, indicate that it has satisfactory properties as a measurement of learning, and scores for about two-thirds of the sample are available on at least one segment of the task. Trouble-shooting is available to examine the impact on predictive validity. Analyses are planned to relate the initial and learning ability measurements to learning.

Spatial Learning Ability: Training versus Practice Effects

Interpreting the effects of the analogue and strategy training as the basis for spatial learning ability requires that a competing hypothesis, namely practice effects, be examined. The purpose of this study was to compare the analogue training and the strategy training conditions to simple practice without feedback.

Method

Materials. The materials were three forms of the spatial folding task. These three forms were also used in the cognitive modeling study with multiple choice items that was described above. In this study, the forms contained 33 items, which were defined by three problems types, with distractor sets nested within problem types. Thus, degrees of rotation and distractor type was varied within all problem types, while the number of surfaces carried was varied only for the spatial items.

Subjects. The subjects were 34 college undergraduates who were participating in the experiment to fulfill a course requirement.

Design. The experiment included three conditions: 1) Analogue Training, 2) Strategy Training and 3) Practice. The conditions were named for the type of training that occurs immediately following the pretest.

The interventions used in the Analogue Training condition and the Strategy Training condition were identical to the interventions in the Lacklund study. Additionally, a Practice intervention unit was developed by presenting the same items that appear in the Analogue Training condition, but without any special instructions, cues or feedback. Comparisons on the first posttest reflect the impact of the three training conditions.

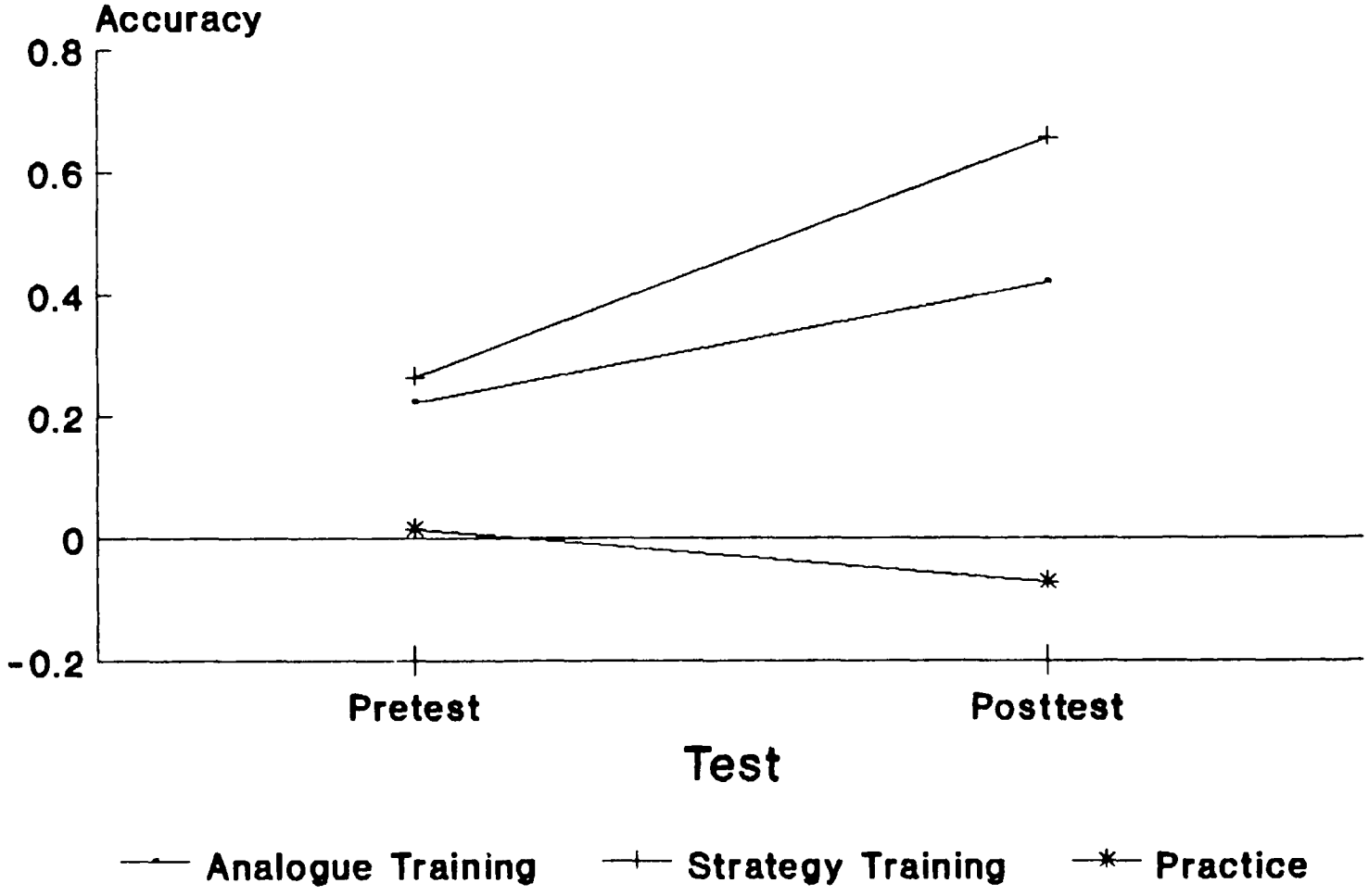
However, some cross-over comparisons were also desired, so that all groups received a second training condition. In the Analogue Training condition, the second intervention was strategy training, which was followed by the second posttest. In the Strategy Training condition, the second intervention was analogue training. In the Practice condition, the second intervention also was analogue training.

Procedure. Subjects were randomly assigned to conditions. All materials were presented on microcomputers in a small laboratory. The subjects were given the standard instructions about the Spatial Learning Ability Test preceding the pretest. The testing session lasted about one and a half hours.

Results

Prior to data analysis, the data were trimmed for subjects with low accuracy rates. Two subjects were eliminated because the first posttest scores were lower than guessing level. Furthermore, two additional subjects, due to an equipment problem, did not reach the second posttest. Because e the sample size was quite small, it was desirable to retain these two subjects. Thus, the current analysis will include only the first pretest.

Figure 8-1 shows the accuracy levels at pretest and posttest for the



Accuracy is ln(P/1-P)

Table 8-1

Analysis of Variance for Accuracy[1]

for Three Dynamic Testing Conditions

at Pretest and Posttest

Effect	df	MS	F	P
Between subjects				
Condition	2	1.30	.83	.44
Error	29	1.55		
Within subjects				
Tests	1	.43	1.57	.22
Condition X Tests	2	.30	1.10	.35
Error	29	.29		

[1] Measured as log odds (ln (P/1-P))

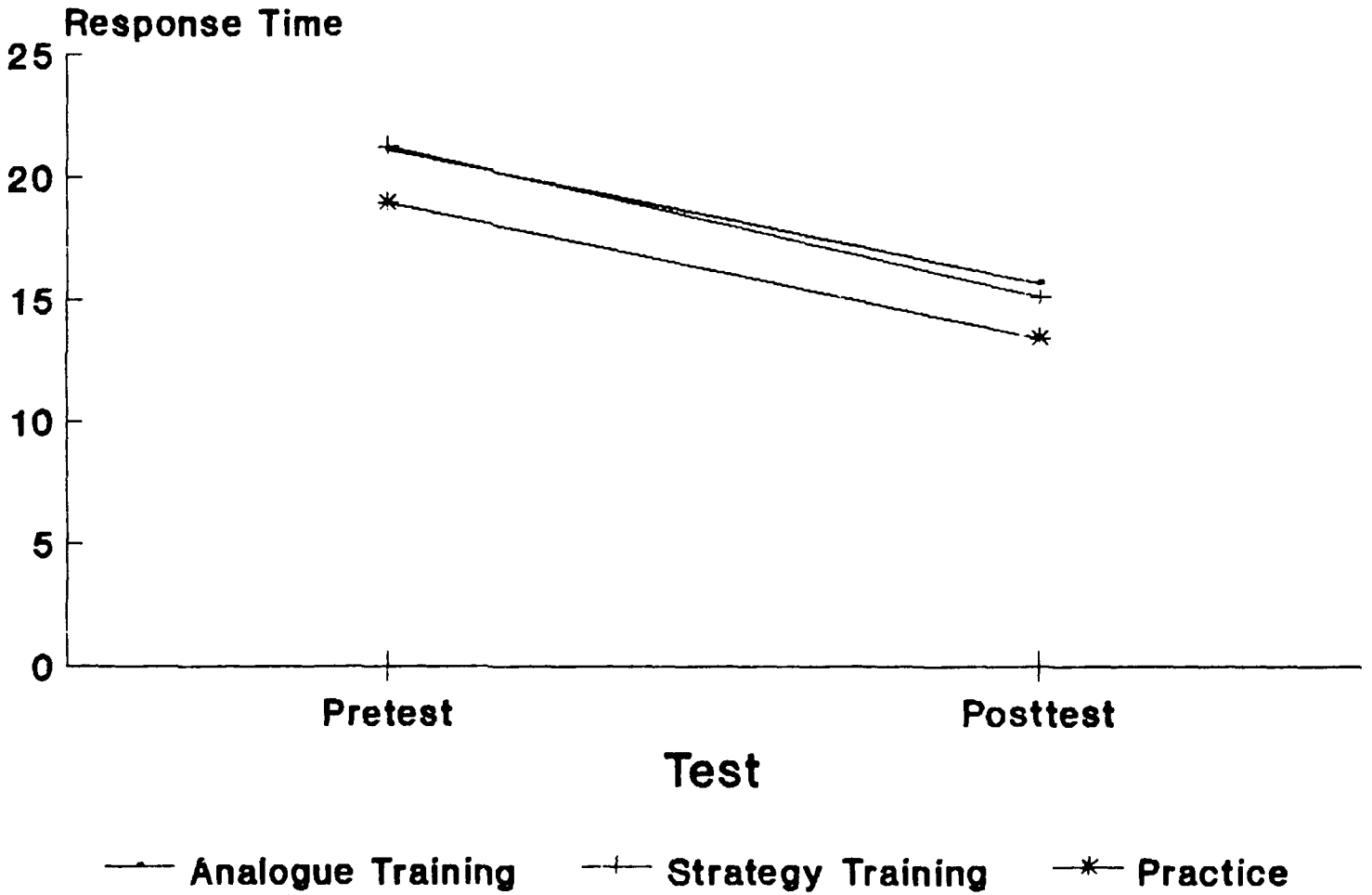

Response Time
25
20
15
10
5
0
Pretest
Posttest
Test
Analogue Training
Strategy Training
Practice

Table 8-2

Analysis of Variance for Total Response Time for Three Dynamic Testing Conditions at Pretest and Posttest

Effect	df	MS	F	P
Between subjects				
Condition	2	27.79	.59	.56
Error	29	46.90		
Within subjects				
Tests	1	517.99	41.71	.000
Condition X Tests	2	.73	.06	.943
Error	29	12.42		

three conditions. It can be seen that accuracy increased for both the Analogue Training condition and the Strategy Training condition, but not for the Practice condition. The analysis of variance results that are presented on Table 8-1 indicate that these trends did not reach significance, however, for the relatively small samples in this study.

Figure 8-2 presents the mean response times at pretest and posttest for the three conditions. It can be seen that response time decreased by about the same amount in all three groups. The analysis of variance that is presented in Table 8-2 shows that the testing occasion effect was highly significant, but neither the condition groups nor the condition by testing occasion interaction reached significance.

Discussion

Although the trends in the data are consistent with disconfirming the hypothesis that accuracy is increased merely by practice, the results did not reach significance. The available sample size for this study was limited due to the unavailability of subjects at the end of the semester. Unfortunately, in studies that attempt to produce changes in tasks for which there are wide individual differences, as in ability problems, larger sample sizes are needed to gain sufficient power to find differences between treatments. Thus, at the time of this report, the sample is being increased to over 100 subjects.

Measuring Mathematical Learning Ability

The previous studies on spatial ability support dynamic testing procedures as providing a method that may provide direct assessments of learning ability. Since only one domain, spatial ability was examined, extending the results to other domains would provide a more general validation for the direct measurement of learning ability.

Mathematical reasoning ability is an ability domain that is potentially interesting for measuring learning ability. Mathematical reasoning is increasingly important as the technological level increases in most occupations. Yet, abilities in this area are often deemed low with respect to the requirements for learning technology and science. An indicator of mathematical learning ability could be useful not only for selection, but also for remedial training in mathematics.

Cognitive Theory and Mathematical Problem Solving

Mayer, Larkin and Kadane (1985) define mathematical ability as consisting of both the structure and operating characteristics of the person's information processing system and the relevant knowledge in long term memory. Mathematics tasks include items such as the following mathematical word problem:

'A roll of plastic 250 meters long costs $26. If it takes a length of 2 1/2 meters of this plastic to cover a certain machine, how much will it cost to buy the exact length of plastic needed to cover 600 such machines?'

Mayer et al (1985) postulate four major steps that are involved in solving math problems. Two of them, problem translation and problem integration, refer to the representation of the problem. The remaining two

steps, problem planning and problem execution, refer to the problem solution process. In the problem representation phase words are translated and integrated into a coherent internal problem representation or schema. Problem solution requires deciding upon a strategy and then executing the operations involved in this strategy.

Each of these four major steps require a different type of knowledge. First, for problem translation linguistic knowledge (one needs to know the language in which the problem is formulated) and factual knowledge about the external world (e.g. knowledge of the metric system) are needed. Second, schematic knowledge, the knowledge of different problem types is required for problem integration. A major source of difficulty in solving mathematical word problems may be situated in this phase, namely the representation of relations between variables. Error analyses of subjects' written protocols (Hall, Kibler, Wenger & Truxaw, 1988; Reed, Dempster & Ettinger, 1985) as well as memory studies in which subjects had to recall problems they had read (Mayer, 1982), revealed that wrongly representing relations between variables constituted a main source of errors. Third, strategic knowledge is required for problem planning and monitoring. Finally, problem execution requires algorithmic knowledge. The four different steps of course are not independent. For example, Mayer (1982) showed that the strategy used to solve a word mathematical problem depended upon the problem representation.

Several classifications of mathematical word problems are possible and have been proposed. Basically, two classification criteria may be used to categorize mathematical word problems: semantic content and formal structure. Semantic content is important, not only in problem

representation, but also in subsequent processing. This is contrary to traditional classroom instruction on problem solving, which focuses on formal problem structure only.

Evidence for the importance of semantic content of problems is given by Hinsley, Hayes & Simon, (1977). They identified 18 categories of mathematical word problems, based on subjects free sortings of mathematical word problems. These categories were all based on content. Furthermore, subjects interviewed by Hinsley et al showed evidence of using these categories to guide their problem solving, (e.g., to distinguish between relevant and irrelevant information). Silver (1981) however showed that the free sortings of mathematical word problems of more mathematically capable subjects were more based on mathematical structure than those of less capable subjects. Also, in a study by Schoenfeld and Hermann (1982) subjects tended to classify more on the basis of mathematical structure after they had attempted to solve the problems.

Hall et al (1988) argued that the classification of problems based on semantic content in subjects' free sortings is an epiphenomenon of the problems sharing the same quantitative substructures. Classifying problems then, based on their quantitative substructures, may be functional, as these provide clues to partial solution strategies. However, the matching errors Reed et al (1985) found, in which the solution of a previous problem was matched to a subsequent problem with similar content but different underlying structure, suggest that recognizing problems having the same quantitative substructures might not be that helpful after all. Reed (1987) also showed that the relative salience of differences in formal structure when subjects judged problem similarity again depended upon the

content of these problems. Also, Reed introduced a factor called 'transparency' to account for the effect of semantic problem content upon the easiness with which subjects are able to recognize isomorphic problems in which problems have the same underlying mathematical structure.

Mayer (1981) incorporated all but one of the Hinsley et al. categories in his extensive taxonomy that included 1079 mathematical word problems. The predominant classification principle in his taxonomy was semantic problem content; that is, problems that shared the same source formula were classified in the same category. The term 'source formula' itself refers to a formal aspect of the problems, however the distinction between different categories rested upon their semantic content, as the source formulas for different problems could be structurally identical but different as to their semantic content. Both below and above the category level, further differentiation between problems was based on the problem's formal structures. This resulted in the differentiation of problem families that included several simple and complex categories; and different templates, with different variations and modifications within categories.

Studies on Performance Modifiability

In contrast to performance on spatial tasks, which is highly modifiable (Embretson, 1987; Regian, Shute & Pellegrino, 1985), performance on mathematical word problems has not proven as modifiable. Studies aimed at improving subjects' performances have yielded mixed results.

Several of these studies (Greeno, 1983; Reed et al, 1985; Reed, 1987) used analogies. For example, Reed (Reed, 1987; Reed et al, 1985) investigated how providing subjects with a solution on a related problem could improve their performance on subsequent problem solving. As Gentner

(1982) describes more generally, an analogy is a structure-mapping between two domains, a known or base domain and a target domain. Such an analogy generally preserves relations among objects in the two domains rather than the attributes of the objects. The structure-mapping between a known domain and a target domain then may be used to elicit the transfer between a known and a target domain, or as Greeno (1983) states, to facilitate the acquisition of (representational) knowledge in a target domain. This is especially the case, as Greeno indicates, if a natural representation in the base domain contains the conceptual entities corresponding to those that are to be acquired in the target domain. For example, mixing water of different temperatures is analogous to mixing liquids containing concentrates of different percentages. Subjects' performances on the latter mathematical word problems may improve if subjects are given this analogy, which was demonstrated by Reed and Evans (1987). An analogy thus may be used to elicit a transfer between two domains.

Greeno (1983) tried to improve subjects' performances, by giving subjects, like the subjects in the above mentioned Embretson study, experience with concrete objects. Greeno compared the performance of four groups of seventh graders on distance problems in a pretest-intervention-posttest design: Two groups were given algorithmic instruction (like the inverse relation between multiplication and division), one control group went to a study hall and one experimental group was given the opportunity to operate on trains on a model railroad track. On the posttest all groups were given the distance-rate-time source formula and some examples. Greeno found that the latter experimental group improved more on the posttest than

Reed and collaborators (Reed, 1987; Reed et al., 1985) investigated

how providing subjects with a solution on a related problem could improve their performance on subsequent problem solving. In a first study, he found that only if subjects were given access to a sufficiently elaborated solution, did their performance improve significantly on equivalent problems (with same structure and same content as the example problem). Performance on similar problems (same content but different structure as the example problems) did not improve. The elaborated solutions subjects were given contained tables, representing the variables of the problems. In the second study, when the complexity level of the example problems was varied, subjects did demonstrate some transfer to the similar problems when the transfer problem was more complex than the example problem. However the problems used in the latter study were structurally more related than the problems from the first mentioned study.

In another study, Reed (1987) also included isomorphic (same structure, different content) problems in addition to equivalent, similar and unrelated problems. Subjects were again presented elaborated solutions on example problems, and were asked: 1) to rate the usefulness of a first given problem in order to solve a second given problem; and 2) to write down the equation of each problem or 3) to match similar concepts in each problem. Each task was performed by a different sample of subjects. These tasks were constructed in order to unconfound the ability to notice an analogy (in 1) and the ability to apply an analogy (in 2 and 3), when told of its usefulness, a distinction that was suggested by Gick and Holyoak (1980). The results showed that these two aspects did differ. For example, for mixture problems isomorphic problems were not rated as more useful than unrelated problems, but they appeared to be useful when

subjects had to write down problem equations. However, the semantic content of problems did play a role too. Work problems for all three tasks in general appeared to be easier than mixture problems.

A study by Hall et al (1988) also demonstrates that subjects do not spontaneously notice analogies. A manipulation of the order in which items were presented, such as two isomorphic problems following each other or two similar problems following each other, did not result in a positive transfer, i.e. transfer between isomorphic problems. Rather, a tendency for a negative transfer between similar problems (i.e., the matching errors mentioned by Reed et al (1985)) was indicated.

Finally, the modifiability of mathematical reasoning on Scholastic Aptitude Test problems also has been examined (Embretson, 1989; Wheeler & Embretson, 1986). Using Mayer, Larkin and Kadane's (1984) taxonomy to classify processes, it was has found that the three types of knowledge, factual/linguistic knowledge, schematic knowledge and strategic knowledge, were associated with significant error rates for young adults. More directly relevant to modifiability, however, performance levels for young adults were significantly increased by supplying cues that changed the cognitive load of these knowledge types. The cues that were supplied were the elaborated solutions that had been obtained from a group of honors students. However, the effect size was not large.

An Experimental on Modifying Mathematical Reasoning

Following Mayer (1985), mathematical ability is considered from the perspective of knowledge required for successful performance on the items of a mathematics task. Within this conceptual framework, the method that Reed et al (1985) used to improve subjects' performances on mathematical

word problems was adapted in the present study. Subjects were provided with a schematic problem representation in the form of a table. This intervention was expected to increase subjects' performances on a posttest as compared to their pretest scores. The posttest included both problems that were structurally equivalent to the intervention problems and far transfer problems that were either similar problems (same content but different structure) or isomorphic problems (same structure but different content).

To determine the effect of this intervention, as compared to the effect of practice alone, the performance of an experimental group was compared to the performance of a control group that solved the problems without receiving the specific instructions. Also, because the Reed et al study showed that providing subjects with an extra aid while they are solving new problems might improve subjects' performances, a second experimental group was included. This experimental group received cues with the posttest item, in addition to the instruction that intervented between the pretest and posttest. Subjects in this cued condition were provided with the source formula for each problem on the posttest.

In the present study two problem types were distinguished, rate problems and part problems. Like Mayer's (1982) categories, these two categories were centered around source formulas. The formulas were 1) RESULTING AMOUNT = ORIGINAL AMOUNT x RATE for the rate problems and 2) TOTAL = PART1 + PART2 + ... + LAST PART for the part problems. Thus, the main classification principle was problem structure.

Within the two main categories, further discrimination between problems was based on both semantic content and structure for the rate

problems (as was noticed by among others Mayer (1981) and Hall et al (1988), the two criteria are correlated) and on structure for the part problems. Different from Mayer's taxonomy thus, problem types involving a kind of rate were taken together to form one main category of rate problems.

Besides the structural basis for the two main categories, our categorization might be justified as being relatively basic, in terms of Greeno's (1983) conceptual distinctions. Reasoning about the parts of a total and finding a total by applying a rate to a time quantity be considered as being very general reasoning procedures.

Another purpose of the study was to build a mental model for mathematical word problem solving. Mathematical word problems are a complex of stimulus features that may influence the several stages of problem solving, as postulated by Mayer (1985). Stimulus factors were hypothesized for each stage except the final computational solution stage. These factors were identified and scored for each mathematical word problem on the pretest.

Method

Subjects. Fifty-four college undergraduates, both males and females, from a large mid-western university participated in the experiment. Subjects were earning credits toward grades in an introductory psychology course. The data of four subjects was lost due to equipment failure. Subjects were randomly assigned to one of three conditions, the control condition, the experimental condition and the cued condition.

Materials and procedure. The tests and the instructions were administered by microcomputers, which recorded both subjects' response

choices and response times. Subjects were also given a booklet. On the first page of the booklet the course of activities was described (i.e. pretest,intervention test and posttest). The remainder of the booklet subjects had to use for scratch paper. The two experimental groups also had to write down their problem tables in it after they had been given the intervention.

The dynamic testing procedure included three tests: a pretest, followed by an intervention test, which included practice problems, and last, a posttest. Within each test, all rate problems were given first, followed by the part problems. All three tests included seven near transfer rate problems and five near transfer part problems. The pretest and the posttest in addition contained three far transfer rate problems and five far transfer part problems. The pretest and the posttest were equated for the far transfer problems and all three tests were equated for the near transfer problems.

In the intervention phase (for the experimental group and the cued group), instructions were given on the two problem types, each time followed by the practice problems of the intervention test. First, instructions were given on the rate problem type. Subjects were given the rate problem source formula and a list of its variants. This was followed by elaborated solutions for the seven rate problems. In these elaborated solutions tables were constructed based on the problem's source formula and these tables were used to solve the problem.

Following these example rate problems, subjects were given the seven practice problems, which were equivalent to the problems explained in the intervention. For each practice problem, subjects were given the source

formula for each problem. Subjects were asked to use this source formula to construct a table for the problems in their booklet. After subjects had entered their final answer on the screen, they were shown the elaborated solution.

The second part of the intervention contained the instruction on the part problem type and its course was similar as for the rate problems. First, subjects were given the part problem source formula (i.e. TOTAL = PART1 + PART2 + ... + LAST PART) and elaborated solutions for five part problems. Then subjects received five equivalent practice problems, for which the source formula was given and subjects were asked to write down the problem's table in their booklet. After subjects had entered their final answer on the screen, the elaborated solution for each problem was shown.

Subjects in the cued condition received, in addition to these instructions, in the posttest the source formula for each problem was shown. Subjects in the control condition in the intervention phase saw the same two sets of problems in the intervention as the two experimental groups, the problems that were explained to the two experimental groups and the practice problems. However, control subjects had to solve all of these problems, without being given any instruction nor the elaborated solutions for these problems.

Subjects in the experimental and cued condition thus had to solve a total of 52 problems: 20 pretest problems, 12 practice problems and 20 posttest problems. Subjects in the control condition had to solve 12 more problems in the intervention test (i.e. those problems that were example problems for the experimental subjects).

For each item subjects were given a maximum of three minutes to solve them. The total task usually took from one and half to two hours.

Results

Prior to data analysis, the data were trimmed for low accuracy levels. Eight subjects were excluded from further analysis, as their performance was below guessing level (an accuracy rate of .20 for the five alternative problems). Including only those subjects that performed significantly better than guessing (i.e. accuracy rates greater than 35 percent correct) would eliminate 17 subjects, which was too many.

Also prior to analysis, the Cronbach's alpha was computed within conditions. The alpha reliabilities of .910, .861 and, .924 were found for the practice condition, the experimental condition and the cued condition, respectively.

Strategy training effects. Accuracy was scored as log-odds (lnP(1-P)) to be comparable to an item response model scale. A Group(3) X Problem Type(2) analysis of covariance was conducted, with pretest scores as the covariate. Group was a between-subjects factor, while problem type was a within-subjects factor. The analysis was performed on both raw scores and log-odds. Table 8-1 presents the mean response times and accuracies over the various sections for the experimental group and the practice group.

The main effect of Group was significant ($F = 3.07$, $p = .038$ for raw scores; $F = 4.10$, $p = .024$ for log-odds). The main effect of Problem type was significant for raw scores ($F = 7.74$, $p = .008$) but not for log-odds ($F = 2.01$, $p = .164$). The Group with Problem type interaction however was significant for both scores ($F = 3.27$, $p = 0.49$ for raw scores; $F = 3.98$, $p = .027$).

Table 8-1

Mean Total Response Times and Accuracy

for Mathematical Problems in Two Groups

	Practice Group		Experimental Group	
Variables	Accuracy	RT	Accuracy	RT
Pretest				
Overall Score	.59	71.57	.55	68.52
Rate Total	.69	65.09	.63	68.69
Near Transfer	.69	70.34	.65	73.16
Far Transfer	.69	52.84	.57	58.24
Part Total	.48	78.04	.48	68.35
Near Transfer	.49	82.03	.54	70.76
Far Transfer	.47	74.05	.41	65.93
Posttest				
Overall Score	.66	50.31	.53	53.25
Rate Total	.79	47.02	.64	54.80
Near Transfer	.84	41.86	.66	53.87
Far Transfer	.67	59.06	.57	56.96
Part Total	.53	53.60	.42	51.70
Near Transfer	.53	51.46	.46	53.59
Far Transfer	.53	55.72	.39	49.80

The main effect of Group was not in the expected direction. It was expected that both experimental conditions would perform significantly better than the control condition, or at least that the experimental condition that in addition to the instruction on problem types received cues on the posttest, would perform significantly better than the control condition. It was found that the control group and the cued group did not significantly differ from each other, but that the experimental group performed significantly worse than the other two groups.

Analyses of covariance conducted on both raw scores and log-odds for near transfer problems only, further clarified these results. For the near transfer problems only, again the main effect of Group was significant both for raw scores ($\underline{F}$ = 3.29, $\underline{p}$ = .048) and for log odds ratios ($\underline{F}$ = 4.23, $\underline{p}$ = .022). The main effect of Group again resulted from the experimental group performing significantly worse than both the control group and the cued group, the two of which did not significantly differ from each other. However, the main effect of Problem type became highly significant ($\underline{F}$ = 14.91, $\underline{p}$ = .000 for raw scores and $\underline{F}$ = 9.16, $\underline{p}$ = .004 for log-odds), but the Group with Problem type interaction disappeared ($\underline{F}$ = 1.12, $\underline{p}$ = .336 for raw scores and $\underline{F}$ = .38, $\underline{p}$ = .688 for log-odds).

Mental models. Five variables were scored on the items to represent the role of factual and linguistic knowledge. These variables were the number of words, Flesch-Kincaid Reading Level, Dale-Chall Word Frequency, if conversions of units were required in the problem and if the problem involved the metric system. Table 8-2 presents the means, standard deviations and correlations of these five variables with response time and accuracy. It can be seen that the number of words, the Flesch-Kincaid

Table 8-2

Descriptive Statistics for Variables

in the Mathematical Models

Knowledge Type	Mean	SD	Correlation RT	Correlation Accuracy
Factual/linguistic				
(1) Number of Words	31.30	9.23	.39*	-.61**
(2) Flesch-Kincaid Reading Level	6.78	3.17	.49*	-.45*
(3) Dale-Chall Word Frequency	7.18	3.27	.30	-.25
(4) Conversions of Units	.15	.36	.41*	.03
(5) Metric System	.20	.41	.03	-.19
Schematic				
(6) Number of Equations	1.85	.75	.14	.12
(7) Relative Definition of Variables Only	.40	.50	.43*	-.54**
(8) Problem Type (Rate=0; Part=1)	.50	.51	.25	-.45*
Strategic				
(9) Transformation Required to Isolate Unknown	.70	.47	.28	-.35

* $p < .05$, one tail

** $p < .01$, one tail

Table 8-3

Correlations Between Variables

in the Mathematical Models

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(1)	1.00								
(2)	.15	1.00							
(3)	-.05	.57	1.00						
(4)	.19	.65	.42	1.00					
(5)	-.04	.38	.54	.14	1.00				
(6)	-.21	.02	-.29	.28	-.24	1.00			
(7)	.43	.02	-.27	-.34	-.15	.17	1.00		
(8)	.50	-.02	-.18	-.42	-.25	-.07	.61	1.00	
(9)	.33	.09	-.13	-.34	-.23	.02	.53	.65	1.00

Table 8-4

Stepwise Regression of Response Time and

Accuracy on Mathematical Model Variables

	Response Time			Accuracy		
Step	Variable	Beta	R	Variable	Beta	R
1	Flesch-Kincaid	.490*	.490*	# of Words	-.609***	.609***
2	Flesch-Kincaid	.480*		Words	-.555**	
	Relative Definition	.416*	.643**	Flesch-Kincaid	-.372*	.712**
3	Flesch-Kincaid	.124		Words	-.397*	
	Relative Definition	.610**		Flesch-Kincaid	-.387*	
	Transformation Required	.542*	.738**	Relative Definition	-.366*	.785**

* $p<.05$

** $p<.01$

*** $p<.001$

Reading Level and the conversion of units had significant correlations with response time. The number of words and the Flesch-Kincaid Reading Level had significant negative correlations with accuracy.

Three variables represented the difficulty of schematic knowledge; the number of equations to be solved simultaneously, the definition of key variables only relative to other variables and the problem type (i.e., rate versus part problems). It can be seen that the relative definition of the variables had a significant correlation with response time, while both the relative definition of variables and the problem type had significant negative correlations with accuracy.

Only one variable was scored for strategic knowledge, the number of transformations required to isolate the unknown. This variable did not correlate significantly with either response time or accuracy.

Table 8-3 presents the intercorrelations of the independent variables, while Table 8-4 presents the results from a stepwise regression of response time and accuracy on the nine variables. It can be seen on Table 8-4 that the sequential steps led to retaining three variables to predict response time, Flesch-Kincaid Reading Level, the relative definition of variables and the number of transformations required to solve the equations. The final multiple correlation was .738 for response time. The stepwise regression led to retaining three variables for predicting accuracy, the number of words, the Flesch-Kincaid Reading Level and the relative definition of variables. The final multiple correlation was .785.

Discussion

The main effect of Problem type indicated that part problems were more difficult than rate problems. The insignificant interaction between Group

and Problem type for the isomorphically equivalent problems indicated that the significant Group with Problem type interaction for the total test had to be situated in the far transfer problems. For the far transfer problems, the residualized gain of the control group appeared to be about equal for both rate and part problems. Both the experimental group and the cued group showed a residualized loss for the rate problems, while on the part problems the residualized loss of the experimental group further increased, contrary to the cued group that showed a residualized gain. In subjects written protocols, evidence was found for 'model based reasoning', especially for the transfer work problems, that were isomorphic to overtake problems.

Factual knowledge does also play an important role. For example, in one problem, although we provided subjects with factual information on the metric system (1 kilometer = 1000 meters) subjects were unable or unwilling to use this information, as an error analysis revealed that the most frequent error was not taking into account the extra meters. In all conditions and for all but one test this option was either the most frequent chosen option or the second most chosen option, following the correct answer. On both the pretest and the practice problems this was the most frequent chosen option (even chosen more than the correct answer). On the posttest the correct answer was chosen either as frequent as this option (control group), or more frequent (cued group) or the correct answer and one other option (where the metric system is wrongly applied) were chosen more frequently (experimental group). Other problems also showed the difficulty subjects had with the metric system.

In general, the results indicated that, for college students,

interventions about problem schema may actually interfere with mathematical problem solving. College students are a mathematically sophisticated group, and perhaps their own schema are more effective than schema they could elaborate after only minimal instruction. However, the cued posttest, in which the problem schema were supplied, did not interfere with performance, and was as effective as simple practice for college student.

These results suggest that a dynamic testing procedure for mathematical reasoning should employ different measurement conditions than for the spatial ability. For spatial ability, the modifiability was measured only with respect to what was supplied by the subjects from the intervention, since no cues were provided on the posttest. The mathematical reasoning results suggest that a posttest without cues will not show the impact of schematic training. So, for mathematical reasoning, at least one posttest measurement should be obtained when structured cues are provided.

The results from the mathematical modeling of response time and accuracy on the mathematical word problems indicated that moderately good prediction was achieved (multiple correlations of .738 and .785, respectively). For both response time and accuracy, the reading level and the relative definition of variables in the problem were important. That is, mathematical word problems became more difficult as the reading level increased and the problem contained definitions of the variables in terms of each other. Additionally, the number of words also significantly decreased accuracy, while the number of required transformations for solving the equations significantly increased response time.

Perhaps the most interesting feature of these results is the important

role of linguistic variables (i.e., reading grade level and number of words) in the difficulty of mathematical word problems. Since the problems in the current study were based on published items from the Scholastic Aptitude Test, the well-known correlation of mathematical problems with verbal abilities is quite understandable.

Calibration of Mathematical Learning Ability Test

A major goal of the project was to calibrate the Mathematical Learning Ability Test (MLAT) on the large sample at Lacklund Air Force Base. The results from the preceding study were important for the design of the stimuli. That is, the results indicated that mathematical reasoning is not highly modifiable for college students, who are relatively skilled in mathematical reasoning. Although it is not clear that the results obtained on college students will apply to a population that varies more widely in mathematical reasoning, some change in the test design seemed to be required. Since performance levels were relatively higher when structured cues accompanied the items, it was decided to add a learning ability that corresponds to performance under structured cues. Thus, learning ability was measured for mathematical reasoning in two ways: 1) the modifiability of latent response potential when structured cues accompany the item and 2) the modifiability of latent response potential on the standard test, after experience with the structured cues. The structured cues would consist of supplying a generalized formula (in words) with each problem. Thus, the structured cues decrease the cognitive load of schematic knowledge.

The calibration study involved measuring mathematical reasoning under three conditions: 1) no cues are presented (i.e., the standard word problem), 2) the problem schema accompanies the problem (e.g., a conceptual version of a formula for the problem), 3) a final test in which no cues are presented.

Mathematical Reasoning

The Mathematical Learning Ability Test (MLAT) measures quantitative

reasoning by performance on word problems that do not require mathematical competency beyond elementary algebra. In fact, most problems can be solved without algebra. The test items on MLAT are similar to disclosed items from the Scholastic Aptitude Test (SAT), except that no MLAT items have tables of data or figures. The MLAT items are similar to SAT items in both syntactic structure and the type of quantitative transformations that are required to solve the problem.

Method

Test design: Cognitive features. Table 9-1 shows three items that appeared on the mathematical reasoning tests. Each item consists of a short paragraph (typically only one sentence) that contains a mathematical problem, and four alternative answers. Subjects are to choose the one correct answer.

The testing materials consisted of three counterbalanced forms, plus some transfer items. Seven different types of problems were presented, as described on Table 9-2. Within each type, the same underlying formula may be used to solve the problem. Three items for each problem type were constructed. The items were equivalent for syntax and the mathematical transformations and combinations that are required, but they varied in content and numerical quantities. Table 9-3 presents three equivalent items for the "discount" problem. It can be seen that all three require solving the formula (Total = Part1 + Part2 + Part3) but contain different content and numerical quantities.

Additionally, two equivalent transfer items were constructed for three problem types, discount, total cost and work problems. The transfer items contained either dramatically different content, or required somewhat

Table 9-1

Three Mathematical Reasoning Problems

A $11.90 book is sold at a 10 per cent discount. What is the total price reduction then in terms of dollars?

a) $.10 b) $.90 c) $1.19 d) $2.58 e) $10.71

DISCOUNT = COST X DISCOUNT RATE

After having washed his scarf, Joe's scarf had shrunk by 20 per cent. If before it was 1.95 meters long, how much shorter is the scarf now? (Meters is written as 'm')

a) 0.20 m b) 0.39 m c) 0.95 m d) 1.18 m e) 1.56 m

REDUCTION = ORIGINAL AMT X RATE

If colored paper costs $0.24 per square meter, how many square meters of colored paper can be bought for $9.00?

a) 36 b) 37 1/2 c) 38 d) 40 e) 45

TOTAL COST = # UNITS X UNIT COST

Table 9-2

Source Formulas for Mathematical Word Problems

Rate Problems

Discount	= Cost X Discount Rate
Reduction	= Original Amt X Rate
Total Cost	= # Units X Unit Cost
Scale Size	= Real Size X Scale Rate
Total Tax	= Cost X Tax Rate
Total Work	= Time X Work Rate
Result	= Original X Rate

Part Problems

Total	= Part1 + Part2 + Part3
Part1	= Total - Part2 - Part3
Part1	= Total - (Rate2 X Total)

Table 9-3

Isomorphically Equivalent Items for Part Problems

A $36 prize is divided among three children. If the first child gets $8 more than the third child and the second child gets $4 more than the third child, how much money does the child who gets most, receive?

a) $4 b) $8 c) $12 d) $16 e) $24

A 40 centimeter ribbon is divided among three children. If the first child's ribbon is 3 centimeters shorter than the second child's ribbon and the third child's ribbon is 2 centimeters shorter than the second child's ribbon, what is the length in centimeters of the shortest piece of ribbon a child received?

a) 11 b) 12 c) 13 d) 14 e) 15

A total of 24 toys are divided among three boys. The second boy gets 1 toy less than the first boy and the third boy gets 5 toys less than the first boy. How many toys did the boy with the least toys get?

a) 5 b) 8 c) 9 d) 10 e) 12

A total of 45 stickers were collected by three friends. If the second child collected 4 stickers more than the first child and the third child collected 2 stickers more than the first child, what is the amount of stickers collected by the child who collected most?

a) 13 b) 14 c) 15 d) 16 e) 17

TOTAL = CHILD1 + CHILD2 + CHILD3

Table 9-4

Original and Two for Transfers Items for a

Work Problem

If a certain machine can produce 75 items in 20 minutes, then how many HOURS will it take to produce 1,350 items at the same rate?

a. 2 b. 6 c. 18 d. 54 e. 67 1/2

Machine A fills 60 boxes per hour and machine B fills 90 boxes per hour. If machine B is turned on 3 hours later than machine A, after how many hours will machine B have filled as many boxes as machine A?

a. 9 b. 6 c. 5 d. 3 e. 2

Machine P prints 120 cards per hour and machine Q prints 150 cards per hour. If machine Q is turned on 5 hours later than machine P, after how many hours will machine Q have printed as many cards as machine P?

a. 25 b. 20 c. 15 d. 9 e. 4

different mathematical combinations. Table 9-4 shows a transfer problem for the "discount" problems.

Problem types were counterbalanced over the pretest, intervention and posttest conditions. Each of the equivalent items for the seven problem types was assigned to a form. The forms were then assigned to one of three groups and three conditions (pretest, cued posttest, final posttest). Thus, every group received every item and every condition, but the exact condition under which a particular item was received varied across groups. Additionally, three transfer items appeared on the pretest and final posttest. These items were also counterbalanced across tests for the three groups.

Intervention conditions. The intervention condition consisted of instruction about the general schemas that appear in the mathematical reasoning pretest. The schemas that are shown on Table 9-2 were described to subjects in detail, and then the schema were illustrated by providing elaborated solutions with sample problems. The elaborated solutions appeared in the form of tables, following Reed et al (1985). Following the instructions, cued practice problems were administered, where the cue consisted of a formula (in words) that represented the appropriate schema for the problem . The intervention is described in greater detail above.

Subjects. The subjects were 582 Air Force recruits who were completing basic training at Lacklund Air Force Base. The recruits had completed approximately three weeks of basic training at the time of the testing.

Procedure. Three condition groups were defined on the basis of a Latin square design in which the three isomorphically equivalent test forms are

assigned to a condition (pretest, cued posttest and final posttest). Subjects were randomly assigned to a condition group.

All subjects received a pretest of ten items, which includes a test form of the seven isomorphically equivalent items plus three transfer items. Then, the intervention is presented, followed by the posttest with cued problems. The final posttest immediately follows the cued posttest. Similar to the pretest, the final posttest also includes three transfer items.

Results

Descriptive statistics. Table 9-5 presents descriptive statistics for raw scores on the isomorphically equivalent items when administered as pretest, cued posttest and final posttest for 582 Air Force recruits. It can be seen that the mean accuracy (log odds scale) is somewhat higher on the cued posttest than on either the pretest or the final posttest. Furthermore, the standard deviation is somewhat larger on the cued posttest. The response times show a similar pattern. The highest mean and standard deviation on response time is also obtained when items are administered under the cued posttest condition. Thus, performance on the cued test involves more processing, presumably to process the cue, and somewhat greater accuracy.

Table 9-5 also shows the test intercorrelations and internal consistencies. These are not expected to be high, since the pretest and final posttest contain only 10 items each, while the cued posttest contains only seven items. The internal consistency of the cued posttest can be compared to the other two tests by applying the Spearman-Brown prophesy formula. The upgraded internal consistency for the cued posttest is .63,

Table 9-5

Descriptive Statistics for Mathematical Learning Ability Test of Lacklund Sample[1]

	Correlations			Cronbach's	Response Time		Log Odds Accuracy	
	(1)	(2)	(3)	Alpha	$\bar{X}$	SD	$\bar{X}$	SD
1) Pretest	1.00			.55	66.57	23.56	.21	1.10
2) Cued Test	.51	1.00		.54	96.28	42.45	.30	1.46
3) Posttest 2	.52	.58	1.00	.59	64.41	30.66	.24	1.13

which suggests that internal consistency is relatively greater on the cued posttest.

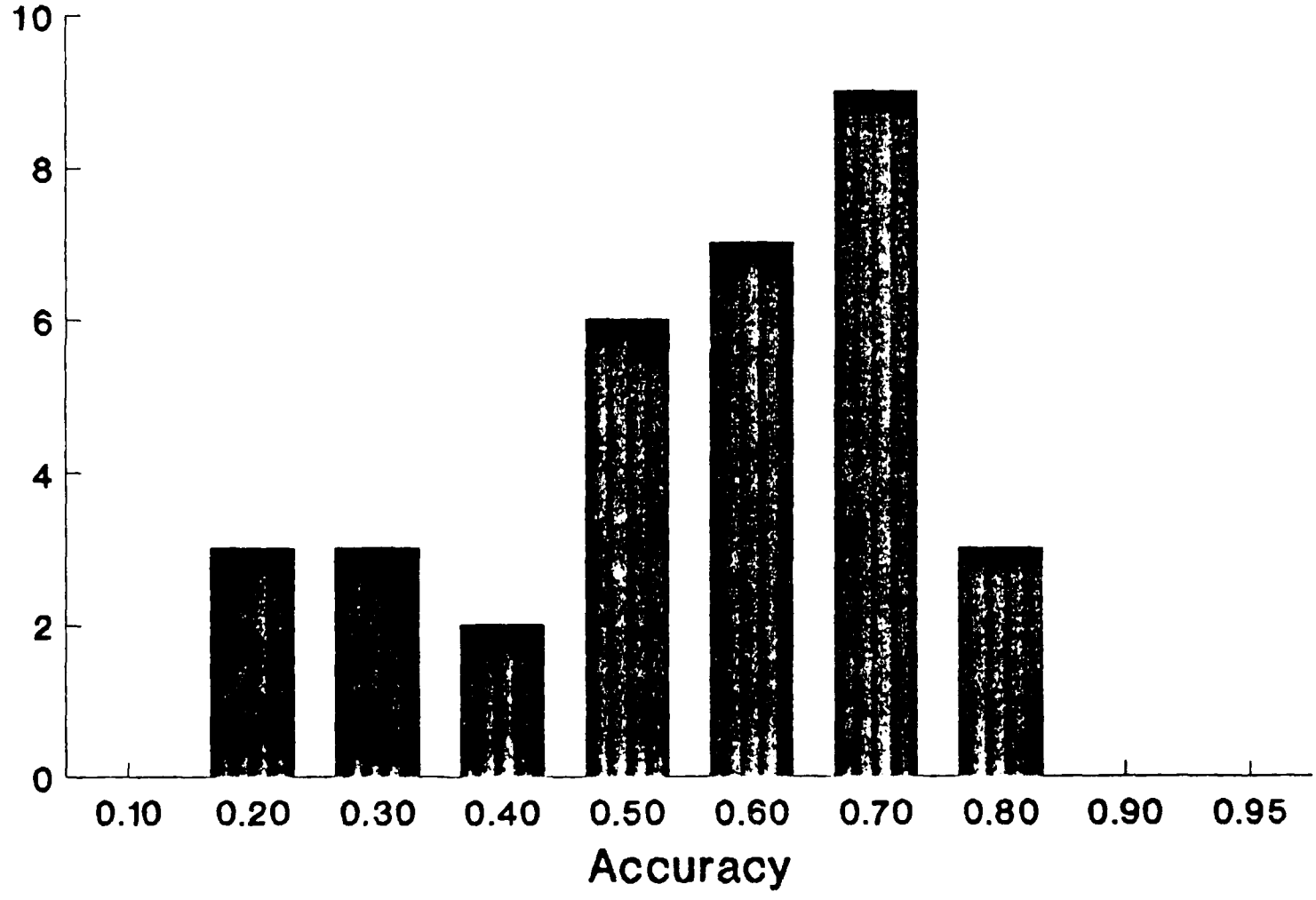
To compare the mathematics tests to SLAT, which has 24 items, the Spearman-Brown prophesy formula was applied to the pretest correlation to posttest correlation (r=.52) and the pretest internal consistency (r= .55). The resulting reliabilities of .72 and .75 indicate that increased test length on MLAT would yield reliabilities that are only somewhat lower than SLAT.

Item characteristics and score distributions. Figure 9-1 presents a distribution of item accuracies on the pretest over the three groups. It can be seen that the test is somewhat easy, since items cluster around a difficulty of .70. However, some extreme items are also included on the test, so that MLAT potentially has some ceiling.

Tables presenting the marginal frequencies of passing items within each group for the pretest, cued posttest and final posttest have been prepared. Additionally, tables presenting the marginal frequencies of scores on the three tests, within each group, have been prepared. These marginals will be used to calibrate the multidimensional item response model for learning and change, and will be included in a forthcoming technical manual.

In addition to parameter calibration, the goodness of fit for the multidimensional item response model for learning and change will also be assessed, using likelihood ratio significance tests, as for SLAT.

For the subject parameters, distributions for initial ability and two learning abilities, cued and uncued, will be prepared.


10
8
6
4
2
0
0.10
0.20
0.30
0.40
0.50
0.60
0.70
0.80
0.90
0.95
Accuracy

Discussion

The results indicate that mathematical reasoning is not highly modifiable by instruction or cues about the problem schema. Although the mathematical reasoning test was somewhat too easy for the population, there was at least some ceiling to observe a marked increase in performance. However, performance levels increased only slightly so that the changes that were observed are not readily interpretable as learning abilities.

Construct Validity for Mathematical Learning Ability

In this section, the results from the Lacklund sample that are relevant to construct validity are described. The data are not completely analyzed at this time.

Method and Results

The materials, procedures and sample is described in the preceding sections.

Impact of training on performance. Table 10-1 presents an analysis of variance on accuracy for the three groups (test form order) and the testing occasion (pretest, cued posttest and final posttest). It can be see that the testing occasion did not reach significance and that a significant group by testing occasion interaction was observed. The significant interaction indicates that the pattern of scores from the pretest to the cued posttest to the final posttest was not uniform between groups.

The results indicate that the (non-significant) main effect for testing occasion should not be interpreted, since the interaction of groups and testing occasion was significant. The groups varied as to which items in the isomorphically equivalent sets were assigned to the pretest, cued

Table 10-1

Analysis of Variance for Mathematical Ability

for Lackland Sample on Log Odds Accuracy

Source	df	MS	F	P
Between Groups	2	5.17	1.64	.20
Error	568	3.16		
Within Groups				
Test	2	1.31	1.83	.16
Test X Group	4	2.16	3.00	.02
Error	1136	.72		

posttest and final posttest. Since groups were randomly assigned, the significant interaction is probably due to differences in difficulty between the isomorphically equivalent items. Estimating the item difficulty parameters in the multidimensional item response model for learning and change will control for any differences in difficulty between the isomorphically equivalent test forms. Thus, the analysis of the effects of intervention must await the final calibration of initial ability and the two learning abilities.

Dimensionality of mathematical ability. The data indicated that the pretest and the cued posttest have patterns of intercorrelations and variances that were consistent with the multidimensional item response model for learning and change. That is, both the variance and the internal consistency increases from the pretest to the cued posttest.

However, the variances and correlations for the final posttest were not consistent with the pattern required by the multidimensional item response model for learning and change. The internal consistency and the variance decreased, rather than increased, from the cued posttest to the final posttest.

Impact of training on mental models. A general impact of training on processing was indicated by the substantial change in response time between the pretest and the cued posttest. However, little difference was observed between the pretest and the final posttest, so it indicates that the changes from the intervention may not show transfer even to equivalent problems.

The more specific impact of the strategy training will be obtained by mathematically modeling both response time and accuracy on the pretest,

cued posttest and final posttest. The variables will be the stimulus features of items that are postulated to influence difficulty in the first three stages of Mayer's (1985) model of mathematical reasoning. That is, the stimulus factors that are postulated to influence factual/linguistic knowledge, schematic knowledge and strategic knowledge will be the independent variables of the model. These variables were used to model response time and accuracy in the experiment that was reported above.

Predictive validity. The Lacklund sample also was administered a computerized training exercise on electronics trouble-shooting, the Logi-Gates. The results from this task provide a criterion measure for which to examine the impact of the mathematical learning abilities on predictive validity. The means and standard deviations for the Logi-Gates were presented with the Spatial Learning Ability Test. A similar analysis is planned for the MLAT as for the SLAT.

Discussion

The results generally indicate that mathematical reasoning performance can be increased somewhat by providing cues with the items. Furthermore, the descriptive statistics indicate that the cued posttest has a pattern of variances and correlations that is consistent with the measurement of learning ability in the context of a multidimensional item response model for learning and change. However, it should be noted that the impact of the cues on performance was not substantial. Further data analyses are needed to determine if the change can be meaningfully described as learning ability.

It is clear from the data, however, that performance on the final posttest will not provide a meaningful basis for measuring learning

ability. Performance levels did not increase and the test variances and correlations are not consistent with a multidimensional model of learning ability. Thus, it appears that transfer to even an equivalent set of problems is difficult to obtain by generalized training on the problem schema.

The main implication of the data, then, is that dynamic testing of mathematical reasoning should include only the cued posttest for measuring learning ability. This approach would have the added advantage of reducing testing time, so that additional items could be added to increase reliability.

GENERAL SUMMARY

A criticism of traditional ability tests is that they are static, rather than dynamic, measures of intelligence. That is, they measure what the person has learned, but not necessarily the capacity to learn. This project developed tests of two learning abilities, spatial learning ability and mathematical learning ability, that are based on cognitive theory. In these tests, which consist of a pretest and two posttests, learning ability is the modifiability of a person's performance under conditions that change the cognitive load of the task, such as strategy training or cues. To solve some psychometric problems in measuring change (i.e., the inequivalencies of raw change at different initial performance levels and the unreliability of change scores), the multidimensional Rasch model for learning and change (Embretson, 1987; 1989a; 1989b) was used to estimate learning abilities. Further, the tests were counterbalanced for the stimulus features that influence processing difficulty to assure cognitive equivalency and to observe the impact of strategy training and cues on the mental models used in the tasks.

Three goals were accomplished for each test: 1) large sample data was obtained to calibrate the tests by the multidimensional Rasch model for learning and change, 2) the construct validity of the learning ability measurements was examined and 3) the cognitive theory underlying the tasks in each test was extended. Although the results on mathematical learning ability were not particularly strong, the measurement of spatial learning ability was strongly supported.

Footnotes

[1] This research was a team effort over one year of funding. The student collaborators on the team were exceptionally dedicated and talented. Included on the team were Machteld Hoskens, James McGill, P. Daniel Smith and Mickey Waxman. Without the particular combination of strengths of these students, the project would have accomplished far less and with less success.

[2] The principal investigator has previously published as Susan E. Whitely.

References

Andersen, E.L. (1973). Conditional inference and multiple choice questionnaires. British Journal of Mathematical and Statistical Psychology, 26, 31-44.

Bereiter, C. (1963). Some persisiting dilemmas in the measurement of change. In C.W. Harris (Ed.) Problems in measuring change. Madison: University of Wisconsin Press.

Budoff, M. (1987). Measures for assessing learning potential. In C. Lidz (Ed.) Dynamic assessment. New York: Guilford Press.

Budoff, M., & Hamilton, J.L. (1974). Learning potential among the moderately and severely mentally retarded. Mental Retardation, 12, 33-36.

Campione, J.C., Brown, A.L., Ferrara, R.A., Jones, R.S., & Steinberg, E. (1985). Breakdowns in flexible use of information: Intelligence-related differences in transfer following equivalent learning performance. Intelligence, 9, 297-315.

Carlson, J.S., & Weidl, K.H. (1979). Toward a differential testing approach: Testing the limits employing the Raven's matrices. Intelligence, 3, 323-344.

Carpenter, P. & Just, M. (1982). Spatial ability: An information-processing approach to psychometrics. In R. Sternberg (Ed.), Advances in the psychology of human intelligence. Hillsdale, NJ: Erlbaum Publishers.

Chase W.G., & Clark, H.H. (1972). Mental operations in the comparison of sentences and pictures. In L.W. Gregg (Ed.) Cognition in learning and memory. New York: Wi' y.

Cooper, L.A. (1976). Individual differences in visual comparison processes. Perception and Psychophysics, 19, 433-444.

Cooper, L.A., & Shepard, R.N. (1973). Chronometric studies of the rotation of mental images. In W.G. Chase (Ed.) Visual information processing. New York: Academic Press.

Dearborne, D.F. (1921). Intelligence and its measurement: A symposium. Journal of Educational Psychology, 12, 123-147 & 195-216.

Egan, D.E. (1980). An Analysis of spatial orientation test performance. Intelligence.

Egan, D., & Gomez, L.M. (1985). Assaying, isolating, and accommodating individual differences in learning a complex skill. R.F. Dillon (Ed.), Individual differences in cognition (Vol. 2). New York: Accademic.

Embretson, S.E. (1989b). A multidimensional item response model for learning processes. Paper presented at the European meeting of the Psychometric Society. Leuven, Belgium, July.

Embretson, S.E. (1987) Improving the measurement of spatial ability by a dynamic testing procedure. Intelligence.

Embretson, S.E. (1989) Diagnostic testing by measuring learning processes: Psychometric considerations for dynamic testing. In A. Lesgold, N. Frederiksen, R. Glaser & M Shafto (Eds). Diagnostic assessment of knowledge and skill acquisition. Hillsdale, NJ: Erlbaum Publishers.

Embretson, S.E. (1987) The psychometrics of dynamic assessment. In C. Lidz (Ed.) Dynamic testing. Beverly Hills: Guilford Press, 1[illegible].

Embretson, S.E. (1985) Test design: Developments in psychology and psychometrics. New York: Academic Press.

Embretson, S.E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S.E. (1984) A general multicomponent latent trait model for response processes. Psychometrika, 49, 175-186.

Feuerstein, R. (1979). The dynamic assessment of retarded performers: The learning potential assessment device, theory, instruments and techniques. Baltimore, MD: University Park Press.

Feuerstein, R. (1980). Instrumental enrichment: An intervention program for cognitive modifiability. Baltimore, MD: University Park Press.

Fischer, G.H. (1973). The linear logistic model as an instrument in educational research. Acta Psychologica, 37, 359-374.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. Cognitive Science, 7, 155-170.

Gick, N.L., & Holyoak, K.S. (1980). Analogical problem solving. Cognitive Psychology, 12, 306-355.

Greeno, J.G. (1983). Conceptual entitites. In D. Gentner & A.L. Stevens (Eds.) Mental models. Hillsdale, NJ: Lawrence Erlbaum Publishers, 227-252.

Hall, R., & Kibler, D., Wenger, E., & Truxaw, C. (in press). Exploring the episodic structure of algebra story problem solving. Cognition & Instruction.

Harris, C. (1963). Problems in measuring change. Madison, WI: University of Wisconsin Press.

Hinsley, D., Hayes, R., & Simon, H. (1977). From words to equations: Meaning and representation in algebra word problems. In P. Carpenter & M. Just (Eds.) Cognitive processes in comprehension. Hillsdale, NJ: Lawrence Erlbaum Publishers, 89-106.

Joreskog, K.G. (1974). Analyzing psychological data by structural analysis of covariance matrices. In D.H. Krantz, R.C. Atkinson, R.D. Luce, & P. Suppes (Eds.) Contemporary developments in mathematical psychology. San Francisco: Freeman.

Just, M. & Carpenter, P. (1985). Cognitive coordinate systems: Accounts of mental rotation and individual differences in spatial ability. Psychological Review.

Lord, F.M. (1963). Elementary models for measuring change. In C.W. Harris (Ed.) Problems in measuring change. Madison: University of Wisconsin Press.

Mayer, R.E. (1982). Memory for algebra story problems. Journal of Educational Psychology, 74, 199-216.

Mayer, R.E. (1981). Frequency norms and structural analysis of algebra story problems into families, categories and templates. Instructional Science, 10, 135-175.

Mayer, R., Larkin, J. & Kadane, P. (1984) A cognitive analysis of mathematical problem solving. In R. J. Sternberg (Ed.), Advances in the psychology of human intelligence: Volume 2, Hillsdale, NJ: Erlbaum Publishers.

Pellegrino, J., Murrow, R.J., & Shute, V.J. (1985). Analyses of spatial aptitude and expertise. In S. Embretson (Ed.) Test design: Developments in psychology and psychometrics. New York: Academic Press.

Reed, S.K. (1987). A structure mapping model for word problems. Journal of Experimental Psychology: Learning, Memory, and Cognition, 13, 124-139.

Reed, S.K., Dempster, A., & Ettinger, M. (1985). Usefulness of analogous solutions for solving word problems. Journal of Experimental Psychology: Learning, Memory, & Cognition, 11, 106-125.

Regian, J.W., & Pellegrino, J.W. (1985). The modifiability of spatial processing skills. Paper presented at the 26th annual meeting of the Psychonomic Society, Boston.

Regian, J.W., Shute, U., & Pellegrino, J.W. (1985, November). The modifiability of spatial processing skills. Paper presented at the 26th annual meeting of the Psychometric Society, Boston.

Resnick, L.B., & Neches, R. (1984). Factors affecting individual differences in learning. In R.J. Sternberg (Ed.) Advances in the psychology of human intelligence, (Vol. 2). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schoenfeld, A.H. (1985). Mathematical problem solving. New York: Academic Press.

Shepard, R.N., & Feng, C. (1972). A chronometric study of mental paper folding. Cognitive Psychology, 3, 228-242.

Shepard, R.N., & Metzler, J. (1971). Mental rotation of three dimensional objects. Science, 171, 701-703.

Silver, E.A. (1981). Recall of mathematical problem information: Solving related problems. Journal of Research in Mathematics Education, 12, 54-64.

Spada, H., & McGaw, B. (1985). The assessment of learning effects with linear logistic test models. In S. Embretson (Ed.), Test design: Developments in psychology and psychometrics. New York: Academic.

Thurstone, L.L., & Thurstone, T.G. (1941). Factorial studies of intelligence. Psychometric Monographs, 2, whole.

Vail, D. (1984). MICROCAT Testing System, Assessment Systems Corporation, St. Paul, Minnesota.

Vygotsky, L.S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Wheeler, A. & Embretson, S.E. (1986) The effects of test anxiety on the cognitive components of mathematical reasoning. Unpublished

Whitely, S.E.[1] (1980) Multicomponent latent trait models for ability tests. Psychometrika, 45, 479-494.

Woodrow, H. (1938). The relationship between abilities and improvement with practice. Journal of Educational Psychology, 29, 215-230.

Woodrow, H. (1946). The ability to learn. Psychological Review, 53, 147-158.